Universidade FUMEC
Faculdade de Ciências Empresariais
Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento

Anotação Semântica Automática do Currículo Lattes Utilizando Linked Open Data

Walison Dias da Silva

Belo Horizonte
2015

Walison Dias da Silva

Anotação Semântica Automática do Currículo Lattes Utilizando Linked Open Data

Projeto de dissertação de mestrado apresentado ao Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento como parte dos requisitos para a obtenção do título de Mestre em Sistemas de Informação e Gestão do Conhecimento. Área de concentração: Gestão de Sistemas de Informação e do Conhecimento. Linha de pesquisa: Tecnologia e Sistemas de Informação.

Orientador: Prof. Dr. Fernando Silva Parreiras

Belo Horizonte

2015

Resumo

A Internet possui documentos de todo tipo que pode ser acessada de qualquer lugar e momento. O conteúdo Web possui uma formatação HTML, desenvolvida especificamente para os seres humanos analisarem. Os sistemas de busca tradicionais são imprecisos na recuperação de informações. O governo utiliza e disponibiliza documentos na Web para que os cidadãos e seus próprios setores organizacionais os utilizem, com isso é importante a existência de ferramentas que apoiem na tarefa da recuperação desses documentos. Como exemplo, podemos citar a Plataforma de Currículos Lattes administrada pelo Cnpq.

A Web semântica possui a finalidade de melhorar e solucionar a questão da recuperação das informações, em que as informações é atribuída seus significados, permitindo que tanto as pessoas quanto as máquinas possam compreender o significado de uma informação. Com a ausência da semântica em nossos documentos da Internet, temos que principalmente os resultados das pesquisas tornam-se menos eficazes, com informações desencontradas e ambíguas.

Diante da importância da Plataforma Lattes na sociedade, esse trabalho propõe como objetivo um arcabouço com os conceitos da Web Semântica para anotar automaticamente o Currículo Lattes por meio das ligações de bases abertas. O problema da pesquisa está baseado em saber quais são os conceitos associados à Web Semântica que podem contribuir para a Anotação Semântica Automática do Currículo Lattes utilizando o Linked Open Data?

Assim através de uma Revisão Sistemática da Literatura o trabalho apresenta conceitos, ferramentas, tecnologias que serão necessárias para o desenvolvimento do framework.

Palavras-chaves: Anotação Semântica. Dados Abertos Interligados. Lattes.

Abstract

Key-words: Semantic Annotation. Linked Open Data. Lattes.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

Cnpq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
FUMEC	Fundação Mineira de Educação e Cultura
IE	Extração de Informação
LOD	*Linked Open Data*
OWL	*Web Ontology Language*
PAS	Plataforma de Anotação Semântica
PLN	Processamento de Linguagem Natural
PRSL	Protocolo de Revisão Sistemática da Literatura
RDF	*Resource Definition Framework*
RSL	Revisão Sistemática da Literatura
SIGC	Sistema de Informação e Gestão do Conhecimento
UML	*Unified Modeling Language*
URL	*Uniform Resource Locators*
URI	*Internationalized Resource Identifier*
W3C	*World Wide Web Consortium*

Sumário

1 Introdução

1.1 Contextualização do Tema

A Internet possui inúmeros tipos de documentos, notícias, informações de todo tipo que pode ser acessada de qualquer lugar e momento. O conteúdo Web possui pouca ou nenhuma estruturação, com uma formatação HTML, onde a maior parte desse conteúdo é projetada para os seres humanos interpretarem, não para programas de computador manipular significativamente. Isso produz nos sistemas de busca tradicionais (utilizam links e palavras chaves) imprecisão e demora na recuperação de informações, isso é, dificultando a realização de pesquisas e retorno de informações mais corretas.

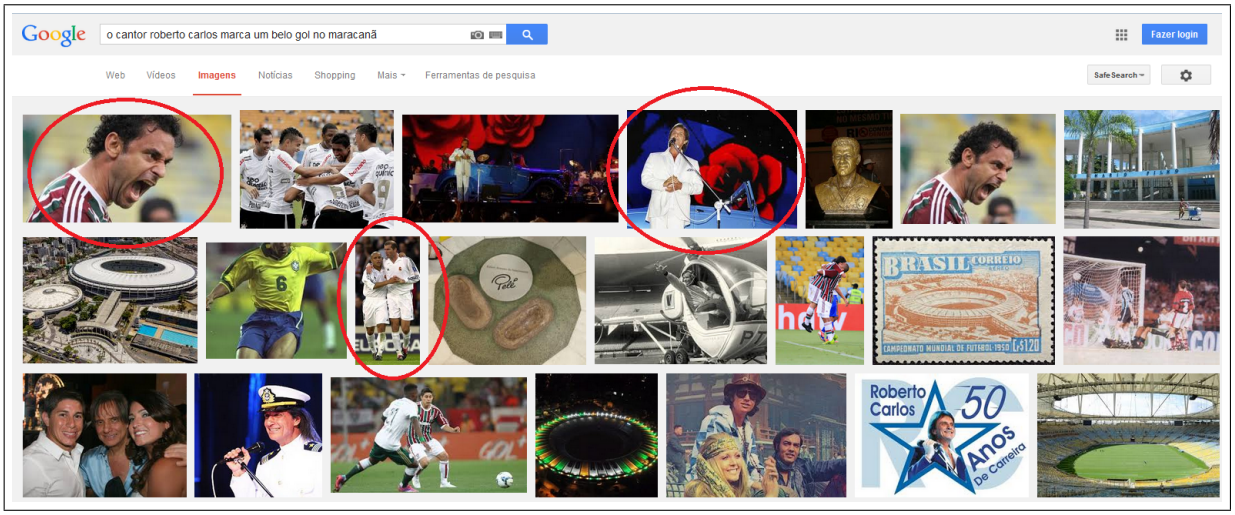

Figura 1 – Resultados da pesquisa no Google sobre Roberto Carlos e seu show no Maracanã

A figura 1 mostra o conjunto de imagens retornadas no Google referente a pesquisa "o cantor roberto carlos marca um belo gol no maracanã". Para as pessoas é notável que estamos nos referindo ao canto Roberto Carlos, mas para a máquina de busca isso não está claramente defino, ocasionando assim resultados imprecisos.

Especificamente, o governo utiliza e disponibiliza sistemas de informações na Web para que os cidadãos e seus próprios setores organizacionais utilizem de suas informações e tanto outras disponíveis na Internet, tais como gráficos estatísticos, legislações e documentos oficiais. Com isso é necessário termos ferramentas e tecnologias que apoie a tarefa de recuperar da Web documentos pertinentes as pessoas e aos setores do próprio Estado com a finalidade de obter resultados atualizados e automaticamente, agregar valor a projetos e políticas públicas. Recuperar informações relevantes no ambiente Web é

uma tarefa trabalhosa e complexa, principalmente quando a mesma envolve características específicas em um determinado domínio (área)(FONTES; CAVALCANTI; MOURA, 2013).

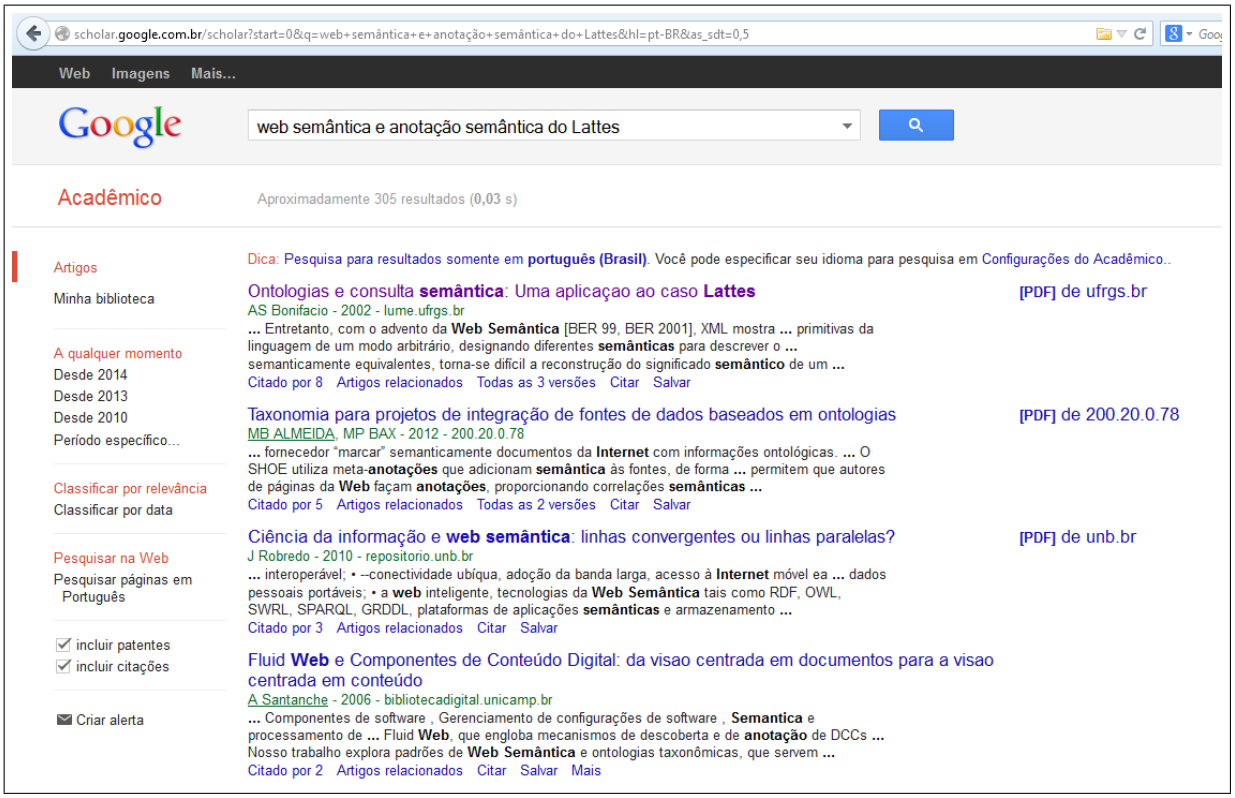

Figura 2 – Resultados da pesquisa no Google sobre web semântica e anotação semântica do Lattes

A imagem 2 refere-se ao conjunto de resultados do Google Acadêmico onde foi realizada a pesquisa "web semântica e anotação semântica do Lattes". Retornando 305 páginas que o usuário deverá navegar em busca do documento que realmente venha introduzir o assunto relacionado ao texto pesquisado, a quantidade e a imprecisão nos resultados ocorre porque os motores de busca utilizam as palavras chaves dos documentos e comparam com as palavras do texto da pesquisa, trazendo em algum momento documentos que tratam de um assunto, em outro momento trazendo documentos relacionados a outra palavra chave, e pior, documentos que não tem ligação com a pesquisa. De qualquer maneira essa avaliação é mais uma tarefa do usuário. O usuário está interessado em encontrar informação onde a relevância dos documentos não pode ser medida através do uso de sistemas de busca por palavras chaves (Keywords) (BONIFACIO, 2002).

Atualmente pode-se recorrer a Plataforma Lattes, que é a base de dados de currículos mantida e administrada pelo CNPq, afim de extrairmos documentos relacionados a esse assunto. Porém temos poucas consultas disponíveis no site da instituição para os cidadãos e outros setores organizacionais sem contar que o acesso a informação está dis-

Figura 3 – Exemplo de arquivo XML disponibilizado na extração de dados na Plataforma Lattes CNPq

ponível em um formato sintático, XML representado na figura 3, tornando o processo de leitura e interpretação das informações pelas pessoas menos legível.

A Web Semântica (WS), proposta em 2001 por (BERNERS-LEE; HENDLER; LASSILA, 2001), como uma extensão da Web atual, onde as informações possuem significado bem definido, permitindo que computadores e pessoas trabalhem em cooperação. A Web semântica surge com o propósito de solucionar o problema de recuperação de dados, interoperabilidade e compartilhamento de conhecimento, em que as informações é atribuída (anotada) seus significados, permitindo que tanto as pessoas quanto as máquinas possam compreender o significado de uma informação. É necessário que o domínio de conhecimento a ser partilhado seja descrito de forma genérica e rica, através de taxonomias e vocabulários específicos, envolvendo conceitos, propriedades e regras de domínio(FONTES; CAVALCANTI; MOURA, 2013). Essa descrição conceitual, também denominada de ontologia, pode ser criada a partir de linguagens lógicas ou ontológicas, como exemplo do RDF (Resouce Description Framework) e OWL (Ontology Web Language), que vai permitir deduzir novas informações sobre um determinado conhecimento, favorecendo a recuperação de informações por agentes inteligentes.

Com a web semântica a Internet passará a funcionar de forma diferente, pois em uma rede de informações, cada item passa a conter o seu significado, o que permite melhores interações com o usuário. Diferente da web tradicional, onde os documentos se relacionam através de links sem significado definido, com essa nova proposta as palavras contém significados que viabiliza sistemas de buscas mais precisos. Assim, não será necessário procurar uma determinada informação em uma série de páginas de resultados

genéricos, será exibido páginas que definem a palavra escolhida.

1.2 Problema

Com a ausência da semântica em nossos documentos da Internet, temos que principalmente os resultados das pesquisas tornam-se menos eficazes, com informações desencontradas e ambíguas. Os motores de buscas tradicionais não compreendem o real significado da intenção da pesquisa, trazendo vários resultados, ficando a tarefa da compreensão dos resultados e a seleção dos documentos para os usuários. Diante das várias áreas que merecem a atenção na resolução deste tipo de problema, escolhe-se trabalhar com uma que seja da sociedade brasileira, que no futuro sirva de base para novas pesquisas e ajude na resolução de problemas para investimento sócio-econômico brasileiro.

Quais são os conceitos associados à Web Semântica que podem contribuir para a Anotação Semântica Automática do Currículo Lattes utilizando o Linked Open Data?

1.3 Justificativa

O desenvolvimento da Internet vive um momento em que conceder significado as palavras, aos conteúdos disponíveis nesse ambiente, é necessário para o seu avanço e crescimento. A Web Semântica estabelece padrões tecnológicos e ferramentas que possibilitarão a criação de novos ambientes informacionais e a efetivação da Web 3.0. Pesquisas na área da anotação semântica são necessárias para solucionar problemas de busca, de localização e de recuperação da informação.

No Brasil a Plataforma Lattes representa a integração de bases de dados de Currículos, de Grupos de pesquisa e de Instituições em um único Sistema de Informações. Sua dimensão atual se estende não só às ações de planejamento, gestão e operacionalização do fomento do CNPq, mas também de outras agências de fomento federais e estaduais, das fundações estaduais de apoio à ciência e tecnologia, das instituições de ensino superior e dos institutos de pesquisa. Além disso, se tornou estratégica não só para as atividades de planejamento e gestão, mas também para a formulação das políticas do Ministério de Ciência e Tecnologia e de outros órgãos governamentais da área de ciência, tecnologia e inovação (CNPQ, 2014).

O Lattes é importante para as Instituições de Ensino, pois o processamento de seus dados e cadastros permite uma fácil visão e avaliação curricular dos docentes e discentes contemplando os seguintes pontos:

1. Estabelecer uma imagem institucional nos sensos;

2. Formação de grupos de trabalho e pesquisa;

3. Avaliar trabalhos de pesquisadores;
4. Diagnosticar o perfil do pesquisador com outros dentro de sua área de atuação;
5. Controle da produção acadêmica docente e discente;
6. Captação de recursos do Estado e agências de fomento;

1.4 Objetivos

1.4.1 Objetivo Geral

Este trabalho tem como objetivo propor um arcabouço com os conceitos da Web Semântica para anotar automaticamente o Currículo Lattes por meio das ligações de bases abertas (Linked Open Data). O trabalho apresentará conceitos, ferramentas, tecnologias que venham enriquecer os dados do Lattes semanticamente, anotar automaticamente por intermédio dos dados interligados. O propósito é permitir que os dados sejam encontrados e indexados na Web, aberto e disponível em formato compreensível por máquina, além de estar disponível para ser reaplicado em outros sistemas e domínios (W3C$_{D}ADOS_{A}BERTOS$, 2014).

1.4.2 Objetivos Específicos

Essa dissertação tem como objetivos específicos:

1. Conceituar e identificar as tecnologias relacionadas com Anotação Semântica e Linked Open Data (LOD).
2. Selecionar e extrair currículos dos docentes da plataforma Lattes/Cnpq.
3. Atribuir significado ao conteúdo do Currículo Lattes com as bases de dados abertas (Anotar o currículo Lattes).

2 Trabalhos Relacionados

Para trabalhos relacionados buscou-se pesquisas que estão relacionadas com estudos sobre a Web Semântica, especificamente na questão de prover semântica a documentos web utilizando um framework para alcançar anotação semântica com LOD. Esse capítulo é organizado primeiramente através de um quadro com as principais características encontradas nos quatro projetos e logo depois uma descrição de cada um dessas pesquisas.

De acordo com a figura 23, observamos que todos os quatro trabalhos possuem como maneira de anotar os documentos o tipo automático, mas somente o trabalho do (FAFALIOS; PAPADAKOS, 2014) é que possui como origem das informações para auxiliar no processo de anotação os dados abertos conectados (LOD). O documento de entrada, aquele que é lido para ser realizado a anotação, diversificou entre os projetos, mas o documento web (Html) foi identificado na pesquisa de (FAFALIOS; PAPADAKOS, 2014) e (NETO, 2009). Para esses trabalhos também temos a característica de extensibilidade, isso é: o sistema que os autores trabalharam no processo de anotação, aceita configurar e ser adaptado a novos domínios e outras ferramentas que auxiliem nesse processo.

Autores	Características dos Trabalhos Relacionados						
	Ferramenta	Doc. Entrada	Origem dos Dados	Extensível	Forma de Anotação	Onde Anota	Estrutura da Anotação
(FAFALIOS; PAPADAKOS, 2014)	Theophrastus	Html, Pdf	Linked Open Data	Sim	Automática	Próprio Doc. Entrada	-
(SANTOS NETO, Gilberto Martins dos, 2009)	Semantic Web Annotation	Html	Específico	Sim	Automática	Gera outro Doc.	OWL
(FONTES; CAVALCANTI; MOURA, 2013)	Autômeta	-	Específico	-	Automática	Próprio Doc. Entrada	RDFa
(ZHANG; CHEN; FENG, 2013)	DBpedia Spotlight	Xml	Dbpedia	-	Automática	Gera outro Doc.	RDFa

Figura 4 – Tabela comparativa entre as características dos trabalhos relacionados

A pesquisa desenvolvida por (FAFALIOS; PAPADAKOS, 2014), *Theophrastus: an demand and real-time automatic annotation and exploration of (web) documents using open linked data*, que suporta a anotação automática de documentos da web por meio de mineração de entidade e fornece serviços de exploração de dados através do conjunto de dados abertos (LOD) em tempo real. Foi apresentado para biólogos da marinha, o sistema atua no domínio de biodiversidade, e tem como objetivo resolver o demorado e

custoso processo de identificação, desambiguar e coletar informações de espécies. É um sistema configurável e adaptável para diferentes domínios de interesse. O Theophrastus não processa RDFa ou Microdados que possam estar em uma página da web, ele apenas identifica as entidades de interesse com base em sua configuração atual e anota as entidades detectadas no próprio documento. O trabalho apresenta algumas ferramentas de extração de entidades baseadas no LOD (DBpedia spotlight, AlchemyAPI, Calais, AIDA e Wikimeta).

O trabalho de (NETO, 2009), *Anotação Semântica De Recursos Web Baseada em Ontologias*, mostrou o desenvolvimento de um método totalmente automático para anotar semanticamente recursos Web, em particular páginas em HTML, visando obter formalmente descrições dos termos existentes nos recursos Web. O projeto foi elaborado em três parte: de extração dos termos dos recursos, mapeamento semântico e anotação dos conceitos identificados, sendo que a ênfase maior é dada ao mapeamento e à anotação semântica, que tem como base a linguagem OWL. Esse projeto gerou um arcabouço de software, Semantic Web Annotation Framework, um sistema de componentes orientado a objetos visando generalização e reutilização. Ele trabalha com duas ontologias, escolhidas para os experimentos nesse trabalho: uma chamada Autos de domínio específico, relacionada a automóveis e outra de domínio genérico, também chamada ontologia de topo, denominada de SUMO. Esse framework é extensível pois sua estrutura permite a utilização de outras ontologias, bastando que elas estejam no padrão OWL DL, também permite a agregação de outros extratores para formatos diferentes de HTML e outras ferramentas de mapeamento semântico que empreguem técnicas diferentes do padrão adotado. Nos trabalhos relacionados na pesquisa de (NETO, 2009), ele destacou as ferramentas Amilcare, SemTag, Seeker, Ont-O-Mat, MnM e WEESA, atentando para essas últimas três ferramentas que há uma desvantagem por serem métodos de anotação semi-automáticos, pois precisam de fases de treinamento e supervisão humana ao contrário do projeto criado por ele. Outra questão que esse autor relata sobre os métodos para a anotação semântica automática SemTag, Seeker e KIM, o principal problema deles é a utilização de uma única ontologia, inviabilizando a aplicação onde o conteúdo dos recursos Web seja de um domínio diferente do domínio da ontologia.

O próximo trabalho, (FONTES; CAVALCANTI; MOURA, 2013), *An Ontology-Based Reasoning Approach for Document Annotation*, apresenta uma proposta para enriquecer automaticamente documentos com anotações semânticas, onde a anotação é realizada de acordo com uma ontologia de domínio. Além de transformar documentos semânticos, também inclui a noção de meta-anotação (anotação sobre anotação). Nessa pesquisa é mencionada algumas outras ferramentas como: Zemanta, Annotea, GATE, SMORE, OpenCalais e GoNTogle. Diferente dessas ferramentas que não exploram intensamente o potencial de inferência de ontologias, ele apresenta como resultado do seu trabalho a ferramenta Autômeta, que tem a funcionalidade de inferir sobre uma ontologia de domí-

nio específico (não trabalha com técnicas de processamento de linguagem natural) e em seguida gerar documentos anotados de forma intrusiva no formato RDFa.

E por último, (ZHANG; CHEN; FENG, 2013), com o texto *Semantic Annotation for Web Services Based on DBpedia*, propõe anotação semântica com base no DBpedia. O enriquecimento é realizado por meio do conjunto de dados abertos interligados (LOD), ontologias do DBpedia. Algumas ferramentas são mencionadas (METEOR-S e ASSAM). Para o seu processo de anotação semântica é utilizado DBpedia Spotlight e o Domj4.

Os trabalhos mencionados acima serão os pilares para o desenvolvimento desta dissertação que possuirá grande semelhança com os primeiros trabalho mencionados. A nossa pesquisa tem a intenção de trabalhar de forma semelhante com os trabalhos de (FAFALIOS; PAPADAKOS, 2014), (NETO, 2009). Assim como na pesquisa desses autores, pode-se reutilizar nesta dissertação a ideia do Extrator de Informação, Mapeador Semântico e a base de dados aberta (Linked Open Data) comentada pelo autor (ZHANG; CHEN; FENG, 2013). Após estudos, aproveitar as ferramentas mencionadas por (ZHANG; CHEN; FENG, 2013) e (ZHANG; CHEN; FENG, 2013) para alcançar o objetivo de anotar os documentos do currículo Lattes.

3 Aderência ao Programa de Pós-Graduação em SIGC da Universidade Fumec

O Curso de Mestrado Profissional em Sistemas de Informação e Gestão do Conhecimento da Universidade FUMEC tem como objetivo geral a geração de novos conhecimentos e a formação de profissionais mestres com habilidades para o desenvolvimento científico, a produção e aplicação prática de conhecimento no campo interdisciplinar de Sistemas de Informação e Gestão do Conhecimento.

Esse curso de pós-graduação stricto senso é organizado sob a área de concentração de Gestão de Sistemas de Informação e do Conhecimento, e possui como linhas de pesquisas: a linha de Tecnologia e Sistema de Informação e a de Gestão da Informação e do Conhecimento.

Essa dissertação desenvolve-se dentro da linha de Tecnologia e Sistema de Informação porque tem como objetivo definir um conjunto de recursos e soluções da computação que permitam o melhor uso da informação. A finalidade dessa linha é a investigação científica relacionada aos processos que podem favorecer o uso da tecnologia no apoio ao gerenciamento dos sistemas de informação.

Por fim esse trabalho apresenta uma temática interdisciplinar, pois envolve disciplinas e conceitos de Gestão do Conhecimento, Sistemas e Tecnologias da Informação conforme identificado na revisão sistemática.

4 Currículo Lattes

De acordo com o CNPq, o Lattes é a base de dados de currículos, instituições e grupos de pesquisa das áreas de Ciência e Tecnologia. Desde os anos 80, já havia interesse por parte dessa instituição de criar um formulário padrão para registrar os currículos dos pesquisadores brasileiros. A partir desse interesse o sistema originou-se com a denominação de Banco de Currículos onde contava com a captação dos dados em papel e em seguida com a digitação no sistema. No início dos anos 90, passa a se chamar BCURR onde a captação dos dados passa a ser em um formulário eletrônico dentro do sistema operacional DOS e em seguida enviado via disquete para ser importado na base de dados. Tempos depois, esse sistema evolui sendo nomeado de Cadastro Nacional de Competências em Ciência e Tecnologia (CNCT) e caracterizado por possuir um formulário eletrônico no ambiente Windows e em seguida os dados eram enviados de forma off line através da Internet. E finalmente, no final dessa mesma década, através dos grupos (CESAR - Centro de Estudos e Sistemas Avançados do Recife - da Universidade Federal de Pernambuco, e o grupo Stela - atual Instituto Stela - da Universidade Federal de Santa Catarina) é desenvolvida uma única versão com a capacidade de integrar as já existentes. Assim, meados do ano de 1999, o CNPq padronizou e lançou o Currículo Lattes para ser o formulário de currículo da esfera do Ministério da Ciência e Tecnologia e CNPq.

A partir de 2001, começou a discussão com relação à abertura e padronização XML com relação ao Lattes. Algumas universidades como UFSC, UNICAMP, UFRJ, USP, UFRGS, UFBA e UFRN solicitaram ao CNPq a abertura tecnológica das informações dessa plataforma, a partir disso, originou a construção da Linguagem de Marcação da Plataforma Lattes (LMPL), sob coordenação da CGINF/CNPQ, sendo os trabalhos de desenvolvimento conduzidos pelo Grupo Stela da UFSC. Mais adiante esse trabalho resultou na formação da Comunidade Virtual LMPL, que definiu o modelo DTD (Data Type Definition) XML do Currículo Lattes. Esse padrão XML foi desenvolvido inicialmente utilizando a linguagem de definição de tipos, DTD (Document Type Definition). Em seguida, com a homologação da linguagem XML Schema pelo Consórcio W3C, a comunidade CONSCIENTIAS-LMPL construiu uma nova estrutura utilizando a linguagem de esquemas para o mesmo padrão XML de Currículo Vitae.

Com isso, tornou-se viável a partir da versão 1.4 do Lattes a abertura da Plataforma, do ponto de vista de conteúdo dos dados, ficando inalterado o acesso técnico às informações, preservando a segurança dos pesquisadores.

O Lattes é um padrão nacional da vida pregressa e atual dos pesquisadores e estudantes. Por sua riqueza de informações e sua crescente confiabilidade e abrangência,

tornou-se um elemento essencial e compulsório à análise de mérito e competência das solicitações de financiamentos na área de ciência e tecnologia. E além de ser um sistema estratégico para as atividades de planejamento e gestão é também utilizado na formulação das políticas do Ministério de Ciência e Tecnologia e de outros órgãos governamentais da área de ciência, tecnologia e inovação.

Qualquer pessoa pode preencher o seu currículo, acessando o site da Plataforma Lattes. Os dados preenchidos são armazenados e disponibilizados publicamente na Internet, tanto em formato HTML quanto em XML. O currículo armazenado na base Lattes recebe um número identificador.

A Plataforma disponibiliza as seguintes funcionalidades:

1. Busca de Currículos: com ela é possível buscar currículos utilizando diversos filtros (como nome, titulação, palavras-chaves etc).

2. Rede de Colaboração: É exibido um grafo no qual os vértices são os pesquisadores e as arestas as colaborações com outros pesquisadores.

3. Painel Lattes: Contém dados estatísticos de toda a base da Plataforma, disposto em forma de gráficos.

Mas, a Plataforma carece de funcionalidades que explorem os dados de um grupo específico de currículos. Certas informações são difíceis de se obter, como por exemplo:

1. Quais professores/pesquisadores publicaram em um determinado ano? Destes, quantas publicações em conferências internacionais? E nacionais?

2. Quais são os professores/pesquisadores de um Departamento ou Universidade? Quais destes possuem registro de publicação? Quais publicaram livros? Quais publicaram capítulos de livros?

3. Quais são os professores/pesquisadores que publicaram em coautoria com um pesquisador/professor?

4. Quais são as teses de doutorado e dissertações de mestrado finalizadas sob orientação de algum professor do grupo nos últimos X anos?

5. Se comparado as publicações com anos anteriores, está havendo um decréscimo ou crescimento no número de publicações de uma determinada Universidade/Faculdade?

6. Os dados de orientações informados por um pesquisador/professor estão condizentes com os informados pelos orientados?

Trabalhos mencionados na próxima seção foram criados com o objetivo de disponibilizar ou facilitar o uso dessas informações.

4.1 Plataforma Lattes e a Web Semântica

Durante o desenvolvimento deste trabalho uma pergunta se fez necessário: "Quais são os trabalhos que associa à Web Semântica com a Plataforma Lattes (PL)?". Diante dessa questão temos:

O trabalho desenvolvido em 2002 por (BONIFACIO, 2002), *Ontologias e consulta semântica: uma aplicação ao caso Lattes*, ele introduz conceitos básicos, uma introdução a novos paradigmas de linguagens e ferramentas que estão dando os primeiros passos em direção a Web Semântica na PL. Um processo de tradução semi-automática dos dados do documento XML gerado pela exportação do Sistema Lattes para o modelo ontológico em DAML+OIL foi apresentado, sendo que a grande contribuição desenvolvida nesse trabalho foi permitir uma melhor compreensão dos conceitos, linguagens e ferramentas que foram apresentadas, com a aplicação prática deles no caso do Currículo Lattes. Como resultado foi elaborado uma proposta de uma ontologia para a plataforma na linguagem DAML+OIL, denominada de OntoLattes.

O Projeto de (CASTAñO, 2008), *Populando ontologias através de informações em HTML - o caso do currículo lattes*, foi um trabalho de dissertação de mestrado onde foi utilizada como fonte de informações os currículos da PL para popular automaticamente uma ontologia (criada por Ailton Sergio Bonifacio e depois convertida de DAML+OIL para OWL por Marcos Yoshinori Nakashima) e utilizá-la principalmente como uma base de dados a ser consultada para geração de relatórios. Todo processo de extração de informações (*wrappers*) foi executado a partir de documentos HTML, com processamento posterior para inserção correta dentro da ontologia, de acordo com sua semântica. Dentro desse processo foi encontrado dificuldades ou problemas como: identificar corretamente os textos dentro dos arquivos originais para que fosse possível mapear a ontologia com a semântica correta dos termos, identificar e retirar as duplicidades de instâncias que se referem a um mesmo objeto. No trabalho foi utilizado duas abordagens diferentes na busca por similaridades e demonstrado suas principais características. Também foi exemplificado de forma superficial uma comparação da criação de consultas em SPARQL, XQuery e SQL.

O trabalho de (GALEGO, 2013), *Extração e Consulta de Informações do Currículo Lattes Baseada em Ontologias*, foi apresentado uma revisão de trabalhos que propuseram a geração de relatórios sumarizados de um grupo de pessoas, alguns até com desenvolvimento de ontologias, dentro do domínio do Lattes. Ele descreve sobre o: OntoLattes que foi a construção de uma ontologia,no formato OWL, para comportar os dados dos currículos dos pesquisadores; sobre o SemanticLattes que realiza as tarefas de importação de currículos e lista de veículos de publicações científicas em duas ontologias (descritas inicialmente em DAML+OIL e em seguida OWL), permitindo consultas às instâncias, ele possui um motor de busca que processa a pergunta em linguagem natural e o software,

por meio de identificação das palavras-chave, reconhece a pergunta e faz a respectiva consulta em SPARQL; sobre o ScriptLattes que é um software que cria relatórios gerenciais obtidos a partir de um conjunto de currículos em formato HTML ou XML, o ScriptLattes não trabalhou com ontologia, as estruturas de dados foram construídas utilizando o conceito de orientação à objetos, ele foi de grande relevância no mundo acadêmico e científico, sendo confundindo muitas vezes com uma ferramenta que foi desenvolvida e disponibilizada pelo CNPq; e por fim o projeto Sucupira, que tem por objetivo a extração de informações da Plataforma Lattes para identificação de redes sociais acadêmicas. Uma das principais funcionalidades deste sistema, Sucupira, é o gerenciamento de uma lista de pesquisadores definida pelo usuário, sendo possível visualizar um mapa contendo o endereço profissional dos pesquisadores, um gráfico sumarizado do número de publicações por ano e tipo, e um grafo relacionando os pesquisadores a outros currículos. Além dessa revisão, (GALEGO, 2013), desenvolve uma ferramenta denominada de Dynamic Lattes, que reutiliza as funcionalidades dos trabalhos citados anteriormente e incorpora novas funcionalidades como a possibilidade de alteração do conteúdo dos dados do relatório sem necessidade de alteração da apresentação, a inclusão do relatório de dados inconsistentes, possibilidade de associar uma orientação a formação de algum membro e resumo da comparação dos dados informados pelo orientador com o orientado.

Uma das questões a ser observada nesse levantamento é com relação ao núcleo das pesquisas mencionadas, conceitos de web semântica e ontologia, não foi encontrado nada especificamente ligado a anotação. Também podemos destacar nesse levantamento a sugestão de trabalho futuro mencionada por esse último autor,(GALEGO, 2013) que é, "Explorar as funcionalidades de Linked Data para que seja possível integração com outras bases de conhecimentos.", que vem de encontro com o núcleo desta pesquisa (anotação semântica com Linked Open Data).

5 RSL - Revisão Sistemática da Literatura

5.1 Fundamento e Conceitos da Web Semântica

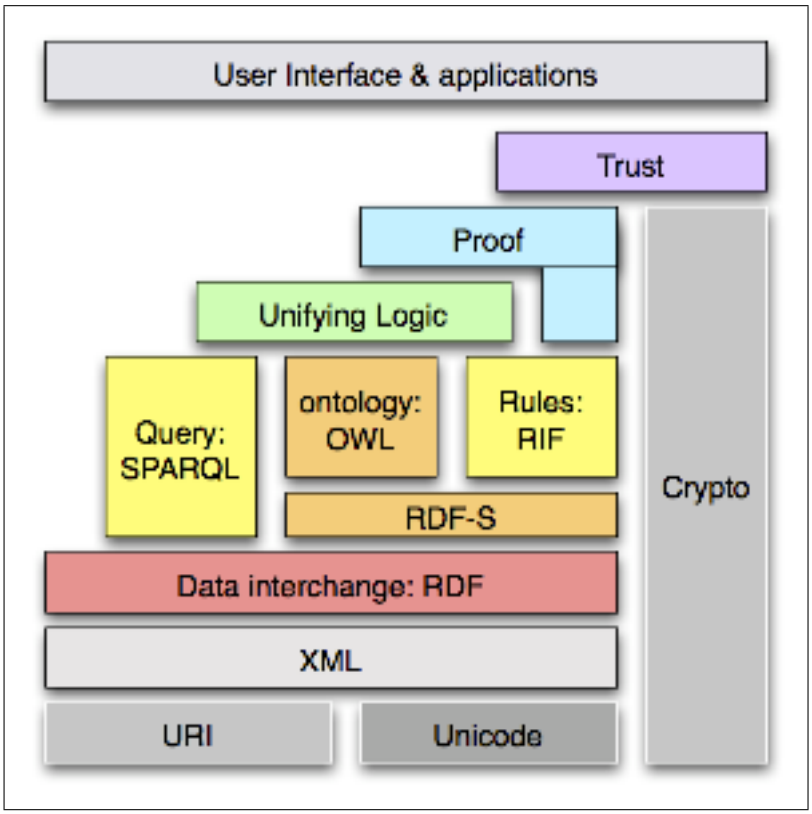

Figura 5 – Arquitetura da Web Semântica.

De acordo com proposta de arquitetura apresentada na figura 5, temos as seguintes tecnologias e camadas relacionada ao projeto Web Semântica:

IRI (International Resource Identifier): é o Identificador Único de Recursos que permite a definição e adoção, de maneira precisa, de nomes aos recursos e seus respectivos endereços na Web. É um padrão para identificar um recurso físico ou abstrato de maneira única e global.

UNICODE: é um padrão de codificação dos caracteres, que diminui consideravelmente a possibilidade de redundâncias dos dados, pois funciona independentemente da plataforma utilizada. Ele fornece uma representação numérica universal e sem ambiguidade para cada caractere de maneira independente da plataforma de software e do idioma.

XML (eXtensible Markup Language): é uma linguagem recomendada pela W3C que permite a criação de documentos que possuem dados estruturados. É uma lingua-

gem que permite a organização dos dados através da definição de elementos e atributos, possibilitando através de regras sintáticas análise e validação de recursos. Fornece a interoperabilidade em relação à sintaxe de descrição de recursos da Web Semântica.

RDF (Resource Description Framework): é um modelo de dados que cria declarações no formato de triplas (sujeito, predicado, objeto), possibilitando a descrição dos recursos por meio de suas propriedades e valores.

RDF Schema: é uma extensão da RDF que permite a definição de esquemas para os vocabulários (termos) utilizados nas declarações. É uma linguagem que permite a construção de ontologias com expressividade e inferência limitadas, pois fornece um conjunto básico de elementos para a modelagem, e poucos desses elementos podem ser utilizados para inferência.

Ontology/OWL(Web Ontology Language): é uma extensão da RDFS que possibilita a inclusão de elementos com maior poder com relação a expressividade e inferência. É uma linguagem para definir e instanciar ontologias na Web. Ela permite a criação de construtos avançados para descrever semântica de declarações RDFS. É baseada em lógicas descritivas atribuindo poder de raciocínio para a Web Semântica.

SPARQL (Protocol and RDF Query Language): uma linguagem de consulta e protocolo de acesso a dados em RDF. Utilizada na recuperação de informações em aplicações da Web Semântica.

Regras/RIF (Regra Interchange Format): é a camada de suporte a regras, RIF é o formato de regras padrão. É importante por exemplo para descrever relações que não podem ser descritas diretamente com OWL. Ela define regras lógicas relacionadas aos recursos informacionais, possibilitando uma espécie de "Introdução Lógica".

Unifying Logic: camada superior que possibilita a incorporação de "Lógicas Avançadas", isso é, responsável pelo raciocínio e inferência a partir de semântica.

Proof: camada responsável por testar a camada de regras e validar as informações. Ela possibilitará a verificação/comprovação da coerência lógica dos recursos, de modo que os aspectos semânticos das informações estejam descritos de maneira consideravelmente adequada, atendendo a todos os requisitos das camadas inferiores.

Trust: é a camada de confiança, local onde se espera garantir que as informações estejam corretas e confiáveis. Camada onde após serem concluídas as informações das camadas anteriores, se determina uma autenticação para que esses dados tornem-se confiáveis.

Interface: é a última camada, onde cumpri-se a interação entre as pessoas e a Web Semântica através de aplicações.

5.1.1 XML - EXtensible Markup Language

A XML é uma linguagem de marcação que trabalha com tags, porém são tags personalizadas (usuários definem suas próprias tags) que permitem a organização e estruturação de dados existentes, possibilitando criar uma marcação específica para praticamente qualquer tipo de informação. O objetivo principal dessa linguagem é descrever informações (foco nos dados). Essa capacidade de descrição é extremamente importante para o armazenamento, recuperação e transmissão dos dados que são estruturados e compartilhados em diferentes sistemas de informação (interoperabilidade de dados).

```
<---XML--->


<Professor>
  <id>12345</id>
  <Nome>Fernando Wagner</Nome>
  <Area>Banco de Dados</Area>
</Professor>
```

Figura 6 – Trecho básico código XML

A figura 6 representa um trecho básico de um documento XML com informações relativa a professor. Os dados estão estruturados de forma que sabemos que o ID "12345" é referente a um professor e seu nome é "Fernando Wagner". O objetivo é estruturar os dados de maneira que os mesmo estejam em condições de sofrerem processamento. É necessário salientarmos que a sintaxe XML para ser escrita deve seguir certas restrições ou regras.

```
<Professor id='12345'>
  <Descricao>
    <Nome>Fernando Wagner</Nome>
    <Area>Banco de Dados</Area>
  </Descricao>
</Professor>
```

Figura 7 – Outra maneira de escrever o trecho básico código XML da figura 6.

Um documento XML pode expressar um conjunto de dados de diferentes maneiras. A figura 7 representa outra forma no qual os dados do professor poderiam ser dispostos no documento XML. Essa característica pode causar divergência de comunicação entre as aplicações e problemas durante o processamento dos dados. Para resolver esse problema são utilizadas linguagens de definição de esquemas que permite especificar como o documento XML deverá ser escrito. Dentre essas linguagens podemos destacar a DTD e XML Schema.

```
1   <?xml version="1.0" encoding="UTF-8"?>
2   <xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
    elementFormDefault="qualified" attributeFormDefault="unqualified">
3   <xs:complexType name="tEndereco">
4      <xs:sequence>
5        <xs:element name="rua" type="xs:string"/>
6        <xs:element name="numero" type="xs:integer"/>
7              <xs:element name="cidade" type="xs:string" minOccurs="1"
         maxOccurs="unbounded"/>
8                 <xs:element name="estado" type="xs:string"
         minOccurs="0" maxOccurs="unbounded"/>
9      </xs:sequence>
10  </xs:complexType>
11  <xs:element name="cliente" type="tEndereco"/>
12  </xs:schema>
```

Figura 8 – Trecho código XML Schema

A figura 8 representa um esquema que especifica quais elementos e atributos são permitidos em um determinado documento XML e como estes devem estar. O DTD é mais antigo e restrito (tipo de dados delimitado a apenas texto), não possuem nenhuma semântica, a validação é apenas sintática, tornando os documentos limitados. Porém o XML Schema resolve em parte esse problema. Ele é baseado na própria linguagem XML, permitindo o reuso do código e ainda disponibiliza maior legibilidade na medida que permite a criação de vocabulários extremamente simples e a definição de tipos de dados como inteiro, binário, entre outros. A W3C recomenda desde 2001 a substituição do DTD pelo XML Schema.

A Extensible Markup Language, XML, carrega a sintaxe das informações, com representação sintática de recursos de maneira independente de plataforma e mesmo com o XML Schema, a XML ainda conta com uma semântica limitada com poucas soluções para o processamento de novos vocabulários. Em resumo, os arquivos XML carregam a sintaxe das informações, mas não a semântica.

Com isso as linguagens RDF e RDF Schema surgem como solução dessas limitações, o que possibilita uma semântica simples relacionadas a identificadores.

5.1.2 RDF - Resource Definition Framework

O XML apesar de ser uma linguagem recomendada pela W3C, ela não possibilita descrever adequadamente a semântica de uma informação. Com isso o modelo de dados, Resource Description Framework (RDF), foi proposto como uma solução para a limitação da XML. Esse modelo de dados é baseado na linguagem XML de modo a expressar o significado das informações, permitindo que essas sejam analisadas sintaticamente, possibilitando interoperabilidade entre aplicações, disponibilizando o conteúdo de forma semântica e compreensível pelas máquinas. Com isso o XML e o RDF tornam-se complementares, a primeira define a estrutura e a segunda permite expressar a semântica

associada aos dados.

O RDF é um padrão de modelo para a troca de dados na Web, criado para situações onde a informação precisa ser processada por aplicativos, ao invés de ser mostrado para pessoas. Esse modelo de dados possibilita a definição de sentenças sobre um recurso.

O modelo de dados RDF é definido como:

· Recursos;

· Literais;

· Propriedades;

· Sentenças.

Podemos entender que um recurso seja "qualquer coisa" sobre a qual se quer expressar uma idéia. Um recurso pode estar relacionado com dados ou com outros recursos através das sentenças. Na terminologia da Web, todos os ítens de interesse são chamados de recursos. O recurso é o mapeamento conceitual para uma entidade ou um conjunto de entidades. Ele é um item com uma característica única, especificado por um URI que é um identificador artificial (não transmite qualquer significado). A URI é semelhante a Uniform Resource Locators (URL), porém tanto pode quanto não pode representar uma página Web.

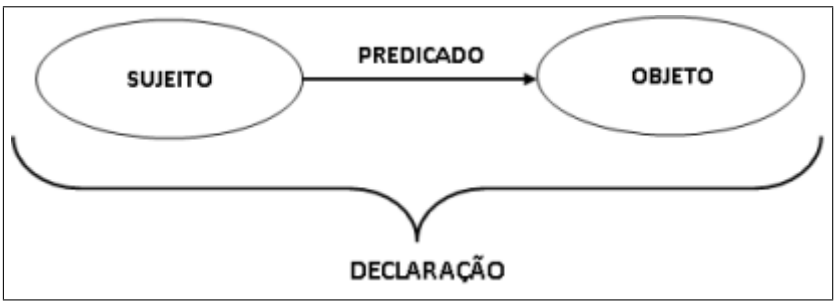

Figura 9 – Modelo gráfico de representação RDF

O relacionamento entre um recurso e um literal é chamado de sentença. De acordo com a figura 9, uma sentença é implementada no formato de triplas (frases formadas com sujeito, predicado e objeto). A sentença relaciona um objeto a um sujeito através de um predicado. O sujeito é uma URI, o objeto pode ser uma URI ou um texto e o predicado define como o sujeito e o objeto se relacionam. O sujeito e o objeto representam um recurso ou itens de interesse em um determinado domínio (BIZER TOM HEATH, 2009).

Um documento RDF pode ser representado de forma abstrata da seguinte maneira:

1. <Bob> <is a> <person>.
2. <Bob> <is a friend of> <Alice>.

3. <Bob> <is born on> <the 4th of July 1990>.
4. <Bob> <is interested in> <the Mona Lisa>.
5. <the Mona Lisa> <was created by> <Leonardo da Vinci>.
6. <the video 'La Joconde à Washington'> <is about> <the Mona Lisa>

Um conjunto de triplas que descrevem informações sobre os recursos envolvidos no domínio de interesse.

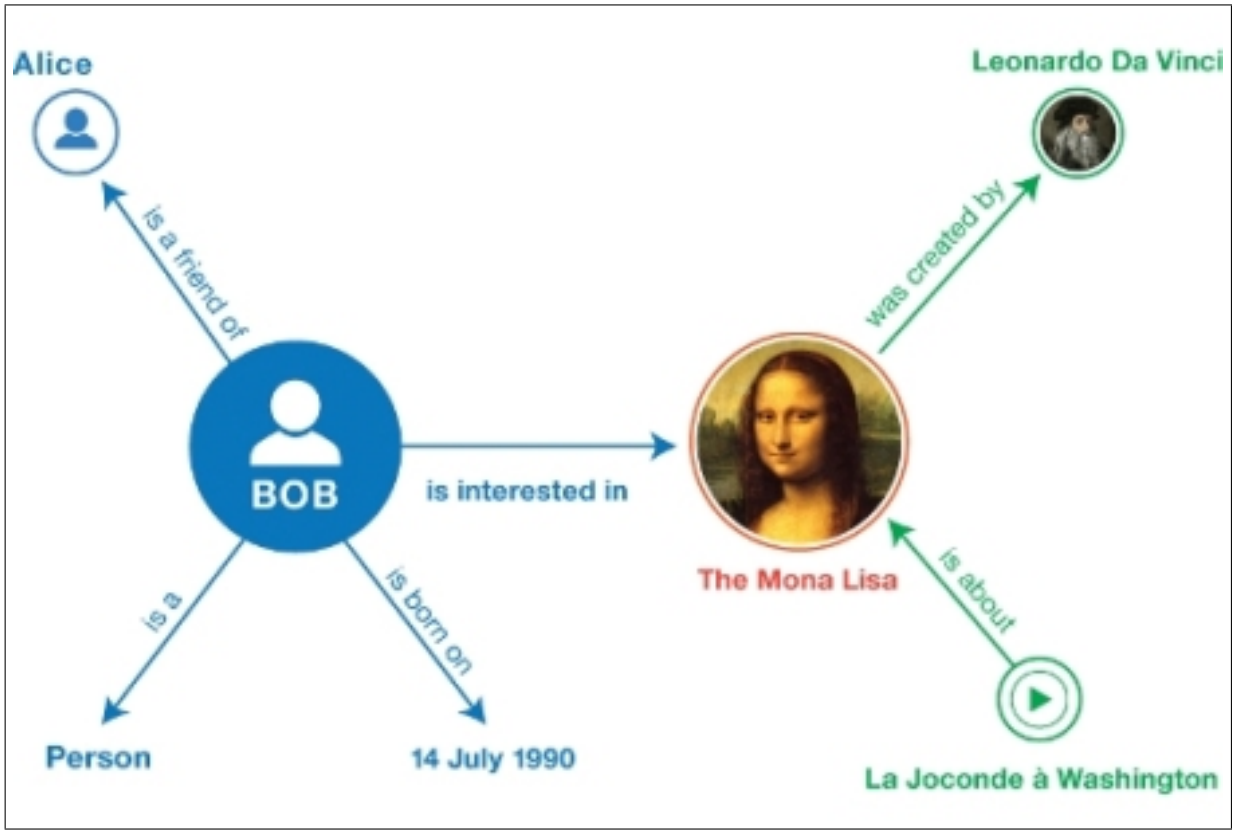

Figura 10 – Representação gráfica de um documento RDF

A figura 10 representa a forma gráfica geral de representar um documento RDF, onde cada recurso e/ou literais existentes são os nós e as propriedades são as arestas.

A figura 11 representa triplas descritas utilizando a sintaxe XML, que habilita o intercâmbio entre máquinas, sem interferência humana através de várias aplicações e serviços.

5.1.3 RDFs - Resource Definition Framework Schema

O RDFs é um complemento para o RDF com o objetivo de oferecer um suporte para a criação de ontologias. O RDFs é uma extensão semântica do RDF, que fornece maneiras

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
         xmlns:dcterms="http://purl.org/dc/terms/"
         xmlns:foaf="http://xmlns.com/foaf/0.1/"
         xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:schema="http://schema.org/">
   <rdf:Description rdf:about="http://example.org/bob#me">
      <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
      <schema:birthDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date">1990-07-04</schema:birthDate>
      <foaf:knows rdf:resource="http://example.org/alice#me"/>
      <foaf:topic_interest rdf:resource="http://www.wikidata.org/entity/Q12418"/>
   </rdf:Description>
   <rdf:Description rdf:about="http://www.wikidata.org/entity/Q12418">
      <dcterms:title>Mona Lisa</dcterms:title>
      <dcterms:creator rdf:resource="http://dbpedia.org/resource/Leonardo_da_Vinci"/>
   </rdf:Description>
   <rdf:Description rdf:about="http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619">
      <dcterms:subject rdf:resource="http://www.wikidata.org/entity/Q12418"/>
   </rdf:Description>
</rdf:RDF>
```

Figura 11 – Representação de Triplas RDF utilizando a sintaxe XML

para descrever grupos de recursos e as relações entre esses recursos. Esses recursos são utilizados para especificar as características de outros recursos, como domínios e faixas de propriedades (W3C$_S$*CHEMA*1.1, 2014).

A ideia principal é unir RDFs + RDF de tal forma que todas as sentenças descritas em RDF obedeçam à semântica descrita no esquema especificado em RDFs. O RDF Vocabulary Description Language ou RDFs é um vocabulário para descrever classes e propriedades dos objetos baseados em RDF com semântica para hierarquias generalizadas dessas propriedades e classes. O RDFs possibilita trabalhar com relacionamentos de abstração de agregação, generalização ou especialização e associação.

Classes descrevem conceitos de um domínio, possibilitando a modelagem do domínio de interesse. As classes são os próprios recursos. Eles são frequentemente identificados por IRIs e podem ser escritos utilizando propriedades de RDF, que é uma relação entre recursos de sujeito e objeto (W3C$_S$*CHEMA*1.1, 2014) .

Construct	Syntactic form	Description
Class (a class)	**C** `rdf:type rdfs:Class`	**C** (a resource) is an RDF class
Property (a class)	**P** `rdf:type rdf:Property`	**P** (a resource) is an RDF property
type (a property)	**I** `rdf:type` **C**	**I** (a resource) is an instance of **C** (a class)
subClassOf (a property)	**C1** `rdfs:subClassOf` **C2**	**C1** (a class) is a subclass of **C2** (a class)
subPropertyOf (a property)	**P1** `rdfs:subPropertyOf` **P2**	**P1** (a property) is a sub-property of **P2** (a property)
domain (a property)	**P** `rdfs:domain` **C**	domain of **P** (a property) is **C** (a class)
range (a property)	**P** `rdfs:range` **C**	range of **P** (a property) is **C** (a class)

Figura 12 – Principais Construtores da Modelagem RDFs.

De acordo com o (W3C$_R$*DF*1.1$_P$*RIMER*, 2014) os principais construtores que permitem especificar formalmente um esquema está sendo demonstrado na imagem 16.

```
1  <?xml version="1.0" encoding="UTF-8" ?>
2  <rdf:RDF xmlns:rdf="http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/#" >
3      <rdfs:Class rdf:ID="Pesquisador"/>
4      <rdfs:Class rdf:ID="Evento"/>
5          <Pesquisador rdf:ID="Walison"/>
6          <Evento rdf:ID="SBBD"/>
7
8      <rdf:Property rdf:ID="Envolve">
9          <rdfs:domain rdf:resource="#Pesquisador"/>
10         <rdfs:range rdf:resource="#Evento"/>
11     </rdf:Property>
12
13     <rdf:Property rdf:ID="Organiza">
14         <rdfs:subPropertyOf rdf:resource="#Envolve"/>
15     </rdf:Property>
16
```

Figura 13 – Declaração de domains, range e subPropertyOf de um RDFS.

A figura 13 é um exemplo de um modelo RDFs: onde na linha 2, destacamos a utilização de *namespaces* visando distinguir o contexto dos elementos utilizados; nas linhas 3-6 apresentam declarações de duas **classes**, *Pesquisador* e *Evento*, juntamente com uma instância de cada uma destas classes. Nas linhas 8-11, é declarada uma **propriedade** *Envolve* que relacionará instâncias de pesquisadores com instâncias de eventos. E nas linhas 13-15, apresentam a declaração de outra **propriedade**, chamada *Organiza*. Os conjuntos *domain* e *range* desta nova propriedade não estão declarados de maneira explicita, mas pode-se inferir que estes têm como valores os conjuntos de pesquisadores e eventos respectivamente, pelo fato de que a propriedade *Organiza* é uma **subpropriedade** de *Envolve*.

5.1.4 OWL - Web Ontology Language

A Linguagem de Ontologia Web (OWL) é utilizada para definir e instanciar ontologias (modelo de dados que representa um conjunto de conceitos dentro de uma área de interesse e os relacionamentos entre essas) na Web com recomendação da W3C. Essa linguagem possibilita que as informações contidas em documentos possam ser interpretadas por humanos e por máquinas, permitindo a interoperabilidade entre aplicações.

Segundo (GALEGO, 2013), OWL foi estabelecido pelo Grupo de Trabalho em Ontologia para Web do W3C como linguagem padrão para construir ontologias para a infraestrutura da Web Semântica. A figura 14 remete a origem da OWL, indicando que ela é o fruto da fusão de duas outras linguagens de descrição: DAML-ONT e OIL. Por isto, inicialmente, era conhecida como DAML-OIL. OWL pode ser entendida como uma extensão do RDF Schema, utilizando dos conceitos já definidos (como rdfs:Class e rdfs:subClass) para suportar uma expressividade mais rica.

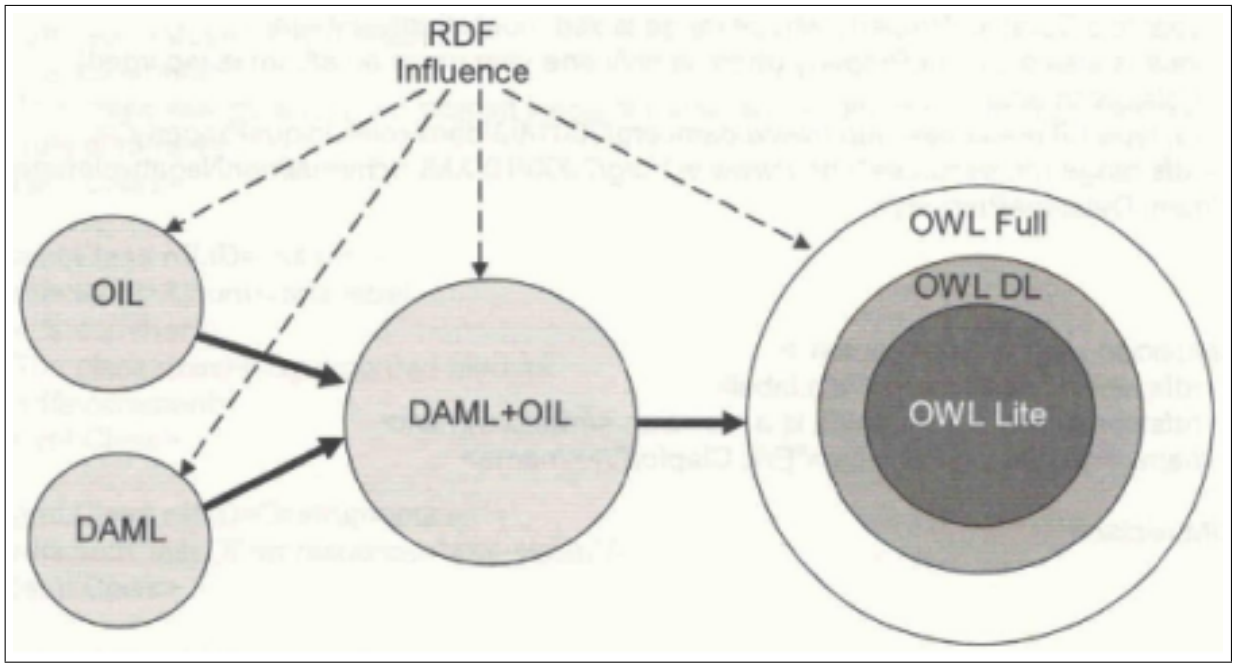

Figura 14 – Influência do RDF na criação da OWL

De acordo com (W3C, 2014), essa linguagem proporciona expressar relacionamentos mais complexos e mais ricos, permitindo, a criação de aplicações com maior poder de inferência ou raciocínio, ela dispõem de tudo que existe no RDFS e outras mais. As sub-línguas para o desenvolvimento de ontologias que são incrementalmente expressivas são:

OWL Lite é adequada para aqueles usuários que necessitam apenas utilizar uma sintaxe simples e que apoie uma hierarquia de classificação e restrições simples. Por exemplo, embora suporte restrições de cardinalidade, ela só permite valores de cardinalidade 0 ou 1. A vantagem é que essa linguagem tem uma menor complexidade formal que OWL DL, sendo de mais fácil entendimento, por parte de usuários, e sendo mais fácil de implementar, pelos desenvolvedores. Ela possui decidibilidade computacional, ou seja, toda a computação terminará em um tempo finito.

OWL DL é utilizada para aqueles usuários que desejam uma expressividade maior do que a oferecida pelo OWL Lite. Baseia-se em lógica descritiva, um fragmento de Lógica de Primeira Ordem, passível portanto de raciocínio automático. Possui completude computacional e decidibilidade. Possui restrições, embora uma classe possa ser subclasse de muitas classes, uma classe não pode ser instância de outra classe. OWL DL é assim chamado por causa da sua correspondência com lógica descritiva, um campo de pesquisa que estuda as lógicas que formam a base formal da OWL.

OWL Full é para aqueles usuários que almejam a máxima expressividade e a liberdade sintática do RDF sem nenhuma garantia computacional. Como exemplo, uma

classe pode ser tratada simultaneamente como uma coleção de indivíduos e como um indivíduo nela própria. Devido a completude da linguagem, é improvável que algum software de inferência venha a ser capaz de suportar completamente cada recurso da OWL Full.

Existe compatibilidade entre as sub-línguas no sentido de que OWL Lite está contido na OWL DL que por sua vez está contida na OWL Full, sendo essa, totalmente compatível com RDF, tanto sintaticamente quanto semanticamente. Geralmente, os documentos OWL disponibilizados na Web são representados pela sintaxe RDF/XML.

Os componentes básicos de uma ontologia OWL são: indivíduos, propriedades e as classes. Estas classes, propriedades e indivíduos podem ser vistos como constituintes atômicos de axiomas e são comumente chamados de entidades.

As **classes** são as entidades principais de uma ontologia e representam os diferentes conceitos de uma ontologia, podendo ser organizadas em hierarquias de superclasse e subclasse, também conhecidas como taxonomias. As classes são conjuntos que contêm os indivíduos e são construídas a partir de descrições, as quais especificam as condições que devem ser satisfeitas por um individuo para que ele possa ser um membro da classe. Subclasses são especializações de suas superclasses.

Propriedades são relações binárias (relações que contém duas coisas) entre indivíduos (instância de uma classe), ou seja, as propriedades ligam dois indivíduos. Na UML é conhecido como relação. Por exemplo, a propriedade hasSibling (temIrmão) pode ligar o indivíduo Pedro ao indivíduo João; ou a propriedade hasChild (temCriança) pode ligar o indivíduo Marcos ao indivíduo Mateus. As Propriedades podem também ser inversas. Por exemplo, a propriedade inversa de hasOwner (temDono) é isOwnedBy (éProprietadeDe). As propriedades podem limitar-se a um valor único: são as Functional Properties (propriedades funcionais). Elas também podem ser Transitive Properties (Propriedades transitivas) ou Symetric Properties (Propriedades Simétricas).

Os **indivíduos** são as instâncias das classes de uma ontologia OWL. Os indivíduos herdam as propriedades das classes de que são membros. É importante mencionar que em OWL dois nomes diferentes podem remeter ao mesmo indivíduo.

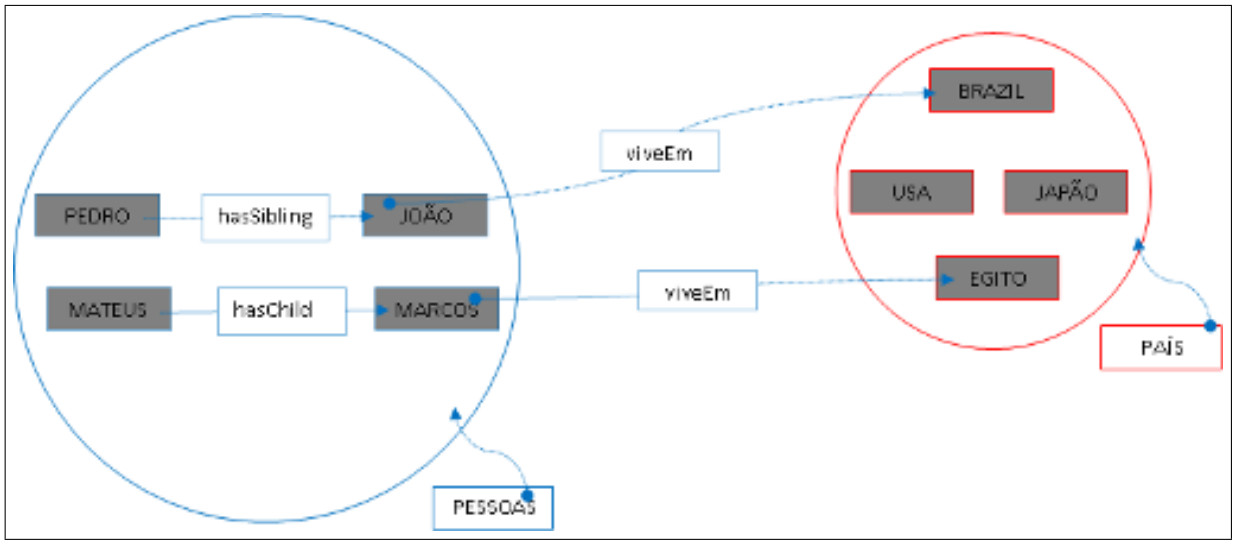

Figura 15 – Exemplo de Relação entre classes e individuos

A figura 15 representa um exemplo de relação entre classes, indivíduo e suas propriedades. Com a OWL é possível combinar diferentes classes e ou propriedades para criar novas classes e propriedades. Essas combinações podem ser complexas e é denominada de expressões. As expressões são a principal razão do poder de expressividade da linguagem OWL. O grupo de trabalho OWL do W3C, em 2009, fizeram uma revisão dessa linguagem e adicionou novos recursos, resultando em uma nova versão da OWL conhecida como OWL 2.

5.1.5 OWL 2

Em 2009, o W3C estendeu as funcionalidades do OWL, definindo uma nova linguagem chamada OWL 2, ficando a primeira versão, de 2004, sendo definida a partir desse momento como OWL 1. De acordo com (W3C, 2015) assim como a OWL 1, OWL 2 é criada para facilitar o desenvolvimento de ontologias e compartilhamento do conhecimento na Web, com o objetivo de expandir a interoperabilidade entre as aplicações de software. A compatibilidade do OWL 2 com OWL 1 é total e completa: Toda ontologia no formato OWL 1 permanece válida para OWL 2, com inferências idênticas em todos os casos práticos.

Com relação ao OWL 1, a OWL 2 adiciona novas funcionalidades. De acordo com (GALEGO, 2013), as novas funcionalidades são para harmonizar a forma sintática (por exemplo, a união disjunta de classes), e outros para oferecer uma nova expressividade:

1. (keys) Definições de chaves para identificar indivíduos de forma única, por exemplo: CPF de uma pessoa, número da placa e estado para um veículo automotivo.

2. (property chains) Propriedades em cadeia: Provê uma forma de definir propriedades a partir de uma composição de outras propriedades. Por exemplo: para definirmos a propriedade TIO, podemos utilizar as propriedades PAI e IRMÃO.

3. (richer datatypes, data ranges) Tipo de dados mais ricos, faixa de dados.

4. (qualified cardinality restrictions) Restrições de cardinalidade qualificadas.

5. (asymmetric, reflexive, and disjoint properties) Propriedades assimétricas, reflexivas e disjuntas.

6. (enhanced annotation capabilities) Estende o uso de anotações, como um comentário ou uma descrição, permitindo aplicar na ontologia, entidades, indivíduos anônimos, axiomas e nas próprias anotações.

Também foi definido novas sub-linguagens ou *profiles*: OWL 2 EL, OWL 2 QL e OWL 2 PL.

OWL 2 EL: É capaz de executar algoritmos em tempo polinomial para todas as tarefas padrões de raciocínio. É particularmente adequado para aplicações com grandes ontologias, nas quais capacidade de expressão pode ser trocada por garantia de performance. **OWL 2 QL**: É capaz de retornar consultas conjuntivas em tempo logarítmico usando tecnologia de banco de dados relacional. É particularmente adequado para aplicações nas quais ontologias relativamente leves são usadas para organizar um extenso número de indivíduos e nas quais é útil ou necessário acessar dados diretamente via consultas relacionais (exemplo: SQL). **OWL 2 RL**: É capaz de executar algoritmos de raciocínio em tempo polinomial usando tecnologias de banco de dados com regras estendidas diretamente sobre as triplas RDF. É particularmente adequado para aplicações que exigem alto grau de raciocínio (por meio de regras de inferência) sem, no entanto, comprometer o poder de expressividade.

5.1.6 Linked Data

Como segmento do desenvolvimento da Web Semântica, o Linked Data (LD) ou dados ligados é um conjunto de boas práticas para publicar e conectar dados estruturados na Web. Estas práticas estão cada vez mais sendo utilizadas levando à criação do que conhecemos como Web of Data(WD). Possibilitando em abrangência global conexão de dados de diversas áreas tais como pessoas, companhias, livros, publicações científicas, filmes, músicas e tantos outros domínios na Web. Tecnicamente, Linked Data diz respeito aos dados disponíveis na Internet que são também compreendidos por máquinas, com significado definido, ligado a outros conjunto de dados externos e que por sua vez, estar conectado a outro conjunto de dados (BIZER TOM HEATH, 2009).

De acordo com (BERNERS-LEE; HENDLER; LASSILA, 2001), o conjunto das melhores práticas a respeito do LD são:

- Usar URI como nome para recursos;
- Usar URI's HTTP para que as pessoas possam encontrar esses nomes;
- Quando alguém procura por uma URI, garantir que informações úteis possam ser obtidas por meio dessas URI, as quais deve estar representadas no formato RDF;
- Incluir links para outros URIs de forma que outros recursos possam ser descobertos;

Como exemplo prático da aplicação dos princípios dos dados ligados (LD), temos o projeto Linking Open Data (LOD), que tem como objetivo principal identificar conjuntos de dados disponíveis sob licenças abertas e convertê-los para RDF de acordo com os princípios Linked Data (BIZER TOM HEATH, 2009). A seguir é exibido esboço do alcance e escala da Web of Data proveniente do projeto LOD, onde cada nó no diagrama representa um conjunto de dados distinto publicado a partir das práticas de Linked Data, em abril de 2014.

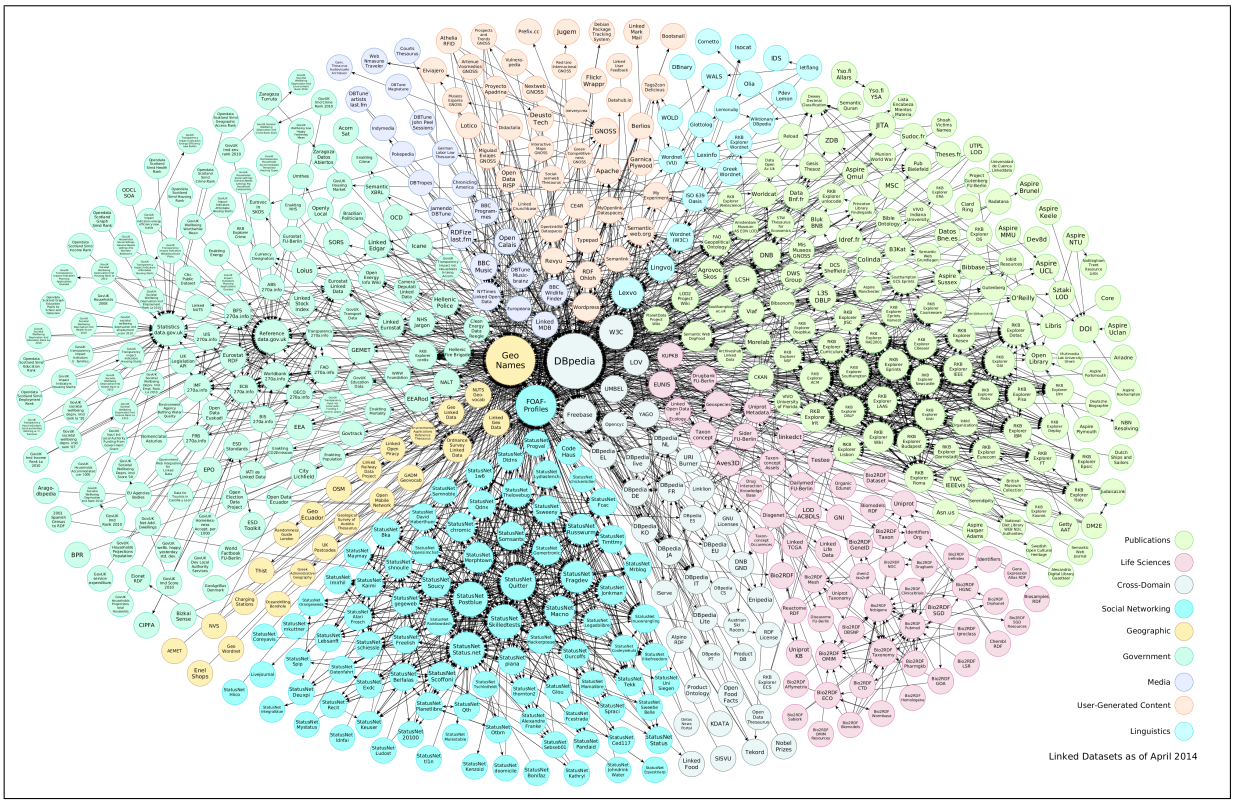

Figura 16 – Linked Datasets as of April 2014

A figura 16 mostra os conjuntos de dados que são publicados como dados ligados à Web, bem como as relações de ligação entre os conjuntos de dados. Ao todo, o dia-

grama contém 570 conjuntos de dados e 2.909 relações de ligação entre os conjuntos de dados(PLANETDATA, 2014).

5.2 Anotação Semântica

Atualmente, a Web Tradicional (WT) é formada por páginas denominas de HTML e a estrutura da Web Semântica (WS) formada por OWL ou RDF. Com o objetivo de criar um documento na WT que seja passível de interpretação humana e que também seja analisado e processado por máquinas e softwares de modo a realizarem pesquisas mais precisas surge proposta da Anotação Semântica (AS)(FONTES; MOURA; CAVALCANTI, 2010).

Web services é considerado em geral uma boa solução para o desenvolvimento de aplicações complexas distribuídas, tais como e-commerce e comunicações. Mas devido a questão de que esses serviços são descritos pelos padões WSDL, os quais geralmente são sintáticos, faltando informações semânticas que proporcionaria respostas mais exatas, a AS é essencial para fornecer os componentes que faltam para o Web service (ZHANG; CHEN; FENG, 2013).

Anotação semântica é uma abordagem para alcançar os conceitos da Web semântica, cuja organização de informações fornece um meio, onde a conexão lógica dos termos estabelece interoperabilidade entre sistemas. Ela é um esquema específico para geração e uso de metadados, habilitando novos métodos de acesso a informação. Uma anotação semântica é uma associação entre as expressões ou termos relevantes de um documento e os conceitos descritos em uma ontologia (BELLOZE et al., 2012).

Segundo (ELLER, 2008) uma anotação semântica de um documento descreve o seu conteúdo pela associação de palavras relevantes do texto e conceitos presentes na ontologia. O resultado de uma anotação **A** é uma tupla (**as, ap, ao, ac**),onde: **as** é o dado anotado; **ao** é a anotação em si; **ap** é o predicado que define o tipo de relacionamento entre o **as** e **ao**; **ac** é o contexto em que a anotação foi feita (OREN et al., 2006).

Para que um documento Web seja bem anotado, é necessária a utilização de múltiplas ontologias ou taxonomias, por isso é essencial uma análise prévia da compatibilidade das ontologias com os domínios dos documentos.

Linguagens específicas para anotação semântica em documentos da internet estão sendo utilizadas, são propostas como as linguagens RDFa, Microformatos e Microdata. Todas se caracterizam por utilizar um conjunto de atributos oriundos de um vocabulário, marcando trechos de um documento HTML ou xHTML, através de triplas semelhantes as utilizadas em RDF (FONTES et al., 2010). Nessa questão destaca-se o RDFa por oferecer recursos mais complexos com *Blank nodes* e o uso de vocabulários arbitrários (FONTES;

MOURA; CAVALCANTI, 2010). A vantagen do RDFa é que ele é recomendado pela W3C e disponibiliza suporte para *Blank nodes* (FONTES; CAVALCANTI; MOURA, 2013). RDFa possui os benefícios do RDF, possuindo recomendação pela W3C para interoperabilidade e legibilidade dos dados, é considerado mais flexível e semanticamente mais rico que os microformatos. Além disso, marcações em RDFa permitem estender funcionalidades, tal como links de URL para as propriedades marcadas. Por isso RDFa é melhor do que Microformatos para anotação de meta-dados (VIRGILIO et al., 2013) (FONTES et al., 2010).

```
<span typeof='ime:Student' about='#Celso'
xmlns:ime='http://ime.eb.br/vocabulary/'>
  Hi! My name is
    <span property='ime:hasName'>
        Celso Fontes</span>
</span>
```

Figura 17 – Exemplo RDFa.

A figura 17 representa um exemplo da aplicação da linguagem RDFa: nos marcadores HTML foram adicionados novos atributos, *typeof* que indica a Classe (Student) do sujeito (Celso) da relação e *property* que indicam respectivamente o predicado ou propriedade *hasName* e a string *Celso Fontes* que representa o objeto. E finalmente, o URI *http://ime.eb.br/vocabulary/* que representa o vocabulário que descreve o contexto em que as descrições do recursos estão definidos (VIRGILIO et al., 2013) e (FONTES et al., 2010). Os novos atributos não interferem na interpretação das informações pelas pessoas e permitem que as informações nos documentos sejam interpretados pelas máquinas.

Sujeito	Predicado	Objeto
about	property	content
src	rel	href
	ver	resource
	typeof	datatype

Figura 18 – Mapa de atributos do RDFa.

Com relação as características e classificação da anotação semântica, temos que ela pode ser do tipo manual onde o usuário faz todo o processo de marcação do documento, selecionando as partes a serem anotada e descrevendo a anotação associada a um termo de uma ontologia. O problema da anotação manual é a grande quantidade de erros, devido a falta de conhecimento ou familiaridade da pessoa que executa a anotação com o domínio, grau de formação, motivação pessoal e complexidade dos esquemas. E ainda, é um processo custoso e que não considera as várias perspectivas de uma fonte de dados (REEVE; HAN, 2005). Segundo (NETO, 2009) a criação de forma manual possui vários problemas como: dificuldade na expressão do conhecimento, alto consumo de tempo e passível de erro.

A outra características e classificação da anotação semântica é a do tipo automática onde uma ferramenta executa a anotação sem a intervenção do usuário, por meio do uso de técnicas como de processamento de linguagem natural (NLP), aprendizado de máquina, extração de informações entre outros, para associar as marcações as expressões da ontologia. A anotação automatizada oferece a escalabilidade necessária para fazer anotações em documentos existentes na Web, e reduz a carga de anotar novos documentos. Outro potencial benefício é a utilização de múltiplas ontologias para anotar um único documento (REEVE; HAN, 2005). E também, temos anotações com suporte manual e automática que são denominadas de híbridas. Outra característica importante é como as anotações são salvas: podendo ser de forma intrusiva quando as marcações ou anotações são armazenadas no documento e de forma não intrusiva quando as anotações são armazenadas em outro arquivo, não modificando o documento original. De acordo com (NETO, 2009), a criação de forma automática possui problemas como: remoção de ambiguidade, obtenção formal de descrições satisfatórias para os conceitos, entre outros.

Na geração semiautomática, é necessária a intervenção do usuário (especialista em gestão do conhecimento) em alguma parte da criação da ontologia ou da anotação.

A tabela 19 representa um resumo das ferramentas de AS com suas características. De acordo com (BELLOZE et al., 2012) explica-se de forma mais detalhada as características de cada uma dessas ferramentas:

1. Annotea: é um projeto da W3C. As anotações desta ferramenta refere-se a comentários, anotações, explicações ou comentários gerais de documentos Web. Ele é parte dos esforços da Web semântica e usa um esquema de anotação baseado em RDF. Os metadados das anotações são armazenados localmente ou em servidores de anotação.

2. Annozilla: é semelhante ao Annotea, porém funciona como um plugin do browser Mozilla Firefox e armazena as anotações em RDF em um servidor. As anotações são destacadas para o usuário mesmo quando a página é recarregada.

3. AutôMeta (Automatic Metadata annotation tool): permite a anotação de um ou mais documentos usando uma ontologia previamente selecionada. As anotações geradas pela ferramenta são armazenadas usando o padrão RDFa.

4. GATE (General Architecture for Text Engineering): é uma ferramenta para aplicações de processamento de linguagem natural. Ele integra um ambiente de desenvolvimento que inclui plugins e outros componentes que permitem tanto a anotação ou extração de informações.

5. Gontogle: é uma ferramenta para anotação e pesquisa. Ele também fornece meios de pesquisar usando uma combinação de busca semântica e as palavras-chave. As

Characteristics of tools. Kind of annotation (A=automatic, H=hybrid, M=manual), Saved annotation (I=intrusive, NI=non-intrusive), Platform (D=desktop, W=web).

Tool	Kind of annotation	Saved annotation	Format of input documents	Format of ontologies	Arbitrary ontology	Platform
Annotea	M	NI	Web documents	-	No	W
Annozilla	M	NI	Web documents	-	No	W
Autômeta	H	I	TXT	N-Triple, RDF, OWL, XML	Yes	D
GATE	H	NI	PDF, TXT, HTML, DOC, ODT	RDF, OWL	Yes	D
GoNTogle	H	NI	PDF, RTF, TXT, DOC, ODT	OWL	Yes	D
KIM	A	NI	HTML	RDF, OWL	No	W
Knowtator	M	NI	PDF, TXT, HTML, DOC, ODT	RDF, OWL, XML	Yes	D
Melita	M	NI	PDF, TXT, HTML, DOC, ODT	OWL	No	D
MnM	M	NI	HTML, TXT	DAML + OIL, RDF	Yes	W
Ontea	A	NI	PDF, TXT, DOC, e-mails, e-mail attachments in HTML	OWL	No	D
RDFaCE	M	I	PDF, TXT, HTML, DOC, ODT	-	No	D
RDFa Editor	A	NI	PDF, TXT, HTML, DOC, ODT	RDF, OWL, XML	Yes	D
Yawas	M	I	Web pages	-	No	W

Figura 19 – Quadro resumo das ferramentas de Anotação Semântica.

anotações são salvas como uma instância no servidor de ontologia e adicionados a uma lista do editor de anotações.

6. KIM: é baseado em uma plataforma Web para pesquisa semântica, anotações de dados e documentos. Ele possui a sua própria ontologia que inclui entidades de interesse geral. O acesso aos recursos da plataforma KIM é realizado através de uma interface Web (KIM Web), que permite métodos tradicionais de pesquisa por palavra-chave ou busca semântica (entidades, padrões).

7. Knowtator: é um plugin do Protégé, e permite um incremento de ontologias para adaptar a aplicação do usuário. A anotação é feita sobre a região do texto selecionado e a especificação da ontologia utilizando as presentes no Protégé.

8. Melita: é uma ferramenta que tem a sua própria ontologia, permitindo aos usuários adicionar os seus resultados para a ontologia, aumentando-a em cada anotação satisfatória.

9. MnM: é uma ferramenta que permite anotações em páginas da Web. Ela utiliza um algoritmo de aprendizado sobre as anotações para posteriormente calcular a precisão e novamente chamar as anotações no corpus. Ele integra um navegador da Web com um editor de ontologia e fornece APIs (Interface de Programação de Aplicativo) para conexão entre servidores de ontologias e ferramentas de extração de informação.

10. ONTEA: utiliza as suas próprias ontologias que estão relacionadas somente a endereços, nomes e e-mails.

11. RDFaCE (RDFa Content Editor): é um plug-in para TinyMCE Javascript Editor WYSIWYG que permite a anotação intrusiva no padrão RDFa. Em vez de ontologias, usa APIs que sugerem os recursos para a anotação. Esses recursos fornecem as URIs para objetos, propriedades e namespaces.

12. RDFa Editor: apresenta-se como uma ferramenta promissora que usa o padrão RDFa para as anotações. Ela permite a utilização arbitrária de ontologias.

13. Yawas: é um plugin de desenvolvido para os browsers Firefox e Google Chrome, onde as anotações são destacadas nas páginas web, mas sem usar alguns recursos semânticos.

No trabalho do (REEVE; HAN, 2005), aborda-se sobre a classificação das Plataforma de Anotação Semântica (PAS), com uma revisão da arquitetura dessas plataformas, suas abordagens e performance. Como demonstrado na figura 20, ele classifica a Plataforma de Anotação Semântica (PAS) em duas categorias: A primeira, **baseada em padrões**: deve ser provido para a plataforma um conjunto inicial de entidades, para que no processo de leitura do corpus sejam encontrados padrões existentes nas entidades. Novas entidades são descobertas, juntamente com os novos padrões. O processo é recursivo até que não haja mais entidades a serem encontradas ou processamento seja interrompido pelo usuário. Nessa classificação também estão anotações geradas a partir de regras geradas manualmente para encontrar entidades em um texto. E a segunda, **baseado em aprendizagem de máquina**: esse possui dois métodos, o probabilístico que utilizar modelos estatísticos para prever e identificar as entidades do texto e o indutivo que reutiliza um

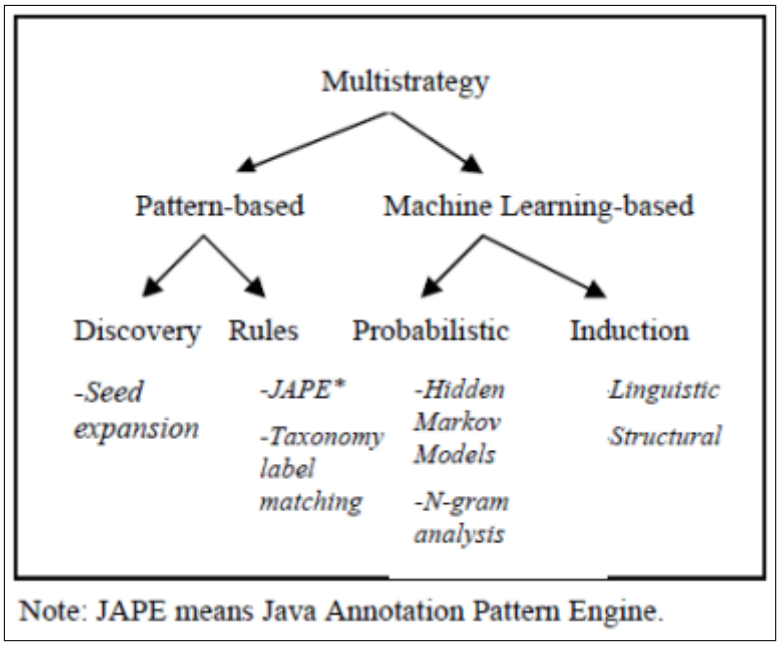

Figura 20 – Classificação das Plataformas de Anotação Semântica

processo de extração de informações para induzir a identificação de entidades. Quanto a arquitetura, está classificada em **não extensiva** que geralmente se concentram em um único domínio, método ou kit de ferramenta; enquanto a **extensiva** permitem que vários componentes do sistema possam ser substituídos ou estendido com outros componentes (isso permite que novos métodos de anotação sejam testados e integrados enquanto reutiliza todos os outros recursos da plataforma).

Nesse seu trabalho é descrito brevemente, através de uma revisão de literatura, algumas PAS afim de mostrar suas arquiteturas e desempenhos medidos empiricamente:

AeroDAML: Possui uma abordagem baseada em padrões, mapeia nomes próprios e as relações comuns com as classes e propriedades correspondentes na ontologia. AeroDAML executa o AeroText, uma API Java, utilizada para acessar o extrator de informações (IE) e mapeá-los em RDF triplica usando uma ontologia como guia.

Armadillo: adota uma abordagem baseada em padrões para encontrar entidades. Ele utiliza o sistema Amilcare como extrator de informações para realizar a identificação de entidades das páginas web de modo indutivo. Uma vez que as origens são encontradas, o padrão de entidades é utilizado para descobrir entidades adicionais.

KIM (Knowledge and Information Management): Contém uma ontologia, uma base de conhecimento, uma anotação semântica, um servidor de indexação e recuperação, bem como front-end para interface com o servidor. O processo de anotação semântica depende de uma ontologia denominada KIMO e de uma base de conhecimento entre domínios. Com o KIMO define-se um conjunto de classes de entidades, relacionamentos e restrições de atributos. Entidades encontradas durante o processo de anotação são compa-

tíveis com seu tipo na ontologia e também com a base de conhecimento, esse mapeamento duplo permite que o processo de extração de informações a seja melhor, fornecendo dicas de desambiguação com base em atributos e relações. A extração de informação é realizada utilizando componentes do kit de ferramentas do GATE.

MnM: disponibiliza um ambiente para anotar manualmente um corpus como exemplo para depois alimentar o sistema de indução que tem como base um algoritmo de processamento de linguagem natural. É uma biblioteca de regras induzidas que pode ser usado para extrair as informações de corpus de textos.

MUSE: foi implementado através do framework GATE e possui a capacidade de realizar reconhecimento e conferencia do nome de entidades. Possui um recurso de processamento para extração de informações (IE) que permite obter precisão semelhante aos sistemas de aprendizado de máquina. Esse sistema é mais sofisticado do que um dicionário de palavras, porque essa não pode fornecer uma lista exaustiva de todos os potenciais-entidades nomeadas, e não pode resolver as entidades ambíguas.

Ont-O-Mat: é um framework de anotação semântica. Seu extrator de informações, Amilcare, é baseado em máquina de aprendizado que exige uma formação corpus de documentos anotados manualmente. Ele também utiliza o ANNIE ("ANearly-New IE system") que é parte do integrante do GATE. O resultado do processamento do Annie é passado para Amilcare, que induz as regras para a IE usando um algoritmo.

SemTag: é uma plataforma abrangente para realizar uma grande escala de anotação de páginas da web. Suas anotações são geradas e armazenadas separadamente do documento de origem. A inteção dessa plataforma é fornecer um repositório público com uma API que permitirá que agentes recuperem a página web de sua fonte e, em seguida, solicitem as anotações separadamente.

Na figura 21, (REEVE; HAN, 2005) demonstra objetivamente os atributos que tem impacto sobre a anotação semântica automatizada. O método utilizado para localizar entidades é o principal determinante no desempenho. Os métodos de aprendizado de máquina, tais como aqueles usados por Amilcare, geralmente possuem um melhor desempenho, embora o sistema MUSE que é baseado em regras, utilizando processamento condicional, mostrou que esse sistema baseado em regras pode igualar o desempenho com sistema baseado em aprendizagem de máquina. Sistemas baseados em regras exigem regras, os sistemas padrão de descoberta requerem um conjunto inicial de dados, sistemas de máquinas de aprendizagem requerem um corpus de treinamento (normalmente anotado), enquanto outros exigem a construção de dicionários para o reconhecimento da entidade nomeada.

A maioria das PAS lidam com um sistema de extração de informação (IE) externa, dos quais a maioria já foram desenvolvidos a partir da comunidade de processamento de

Platform	Method	Machine Learning	Manual Rules	Bootstrap Ontology
AeroDAML [14]	Rule	N	Y	WordNet
Armadillo [10]	Pattern Discovery	N	Y	User
KIM [18]	Rule	N	Y	KIMO
MnM [21]	Wrapper Induction	Y	N	KMi
MUSE [16]	Rule	N	Y	User
Ont-O-Mat: Amilcare [12]	Wrapper Induction	Y	N	User
Ont-O-Mat: PANKOW [5]	Pattern Discovery	N	N	User
SemTag [9]	Rule	N	N	TAP

Figura 21 – Resumo das Características das Plataformas de Anotação Semântica

linguagem natural (PLN). Alguns sistemas de IE dispõe de serviços adicionais, tais como reconhecimento de entidades nomeadas, IE por regras de indução usando a máquina de aprendizagem e encontram relações de identidade entre as entidades em texto (co-referenciação).

Com relação a performance dessas ferramentas, nas suas observações e de acordo com a figura 22, é possível observar que o melhor desempenho foi realizado pela PAS MnM que é baseada em aprendizagem de máquina. A PAS baseada em padrões com pior desempenho foi o MUSE. E a pior de todas as PAS foi o Ont-O-Mat. As medidas padrão de *Precision, Recall*, e *F-measure*, foram obtidas a partir do campo recuperação da informação, utilizados pelos autores das PAS na determinação da eficácia de anotação.

(REEVE; HAN, 2005) conclui a sua pesquisa afirmando que as Plataformas de anotação semântica (PAS) podem ser distinguidas principalmente pelo seu método de anotação, sendo esse o componente que tem o maior relevância sobre a eficácia de uma anotação semântica. Os algoritmos de aprendizado de máquina, possuem um desempenho mais eficaz do que os métodos baseados em padrões, mas foi demonstrado na pesquisa que um sistema baseado em regras utilizando o processamento condicional pode executar tão bem quanto um sistema de aprendizagem de máquina. A contínua evolução das PAS e a ampliação de novos recursos afim de proporcionar uma anotação melhor é fundamental para a realização da Web Semântica.

Framework	Precision	Recall	F-Measure
Armadillo	91.0	74.0	87.0
KIM	86.0	82.0	84.0
MnM	95.0	90.0	n/a
MUSE	93.5	92.3	92.9
Ont-O-Mat: PANKOW	65.0	28.2	24.9
SemTag	82.0	n/a	n/a

Figura 22 – Performance das Plataformas de Anotação Semântica

5.3 Planejamento

O objetivo do "Planejamento"na RSL é construir um protocolo de avaliação, identificando a necessidade da revisão e com quais critérios de busca serão utilizados.

É necessário novos estudos que permitam avanços no setor anotação semântica, principalmente no quesito anotação automática de documentos Web. Fica definido nesta seção que a revisão da literatura tem como proposição responder as seguintes questões:

1. : Qual ou quais componentes/módulos que são necessários para montar uma arquitetura de anotação semântica com Linked Open Data?
2. : Dos componentes/módulos encontrados quais são os mais utilizados?
3. : Dos componentes/módulos encontrados, qual(is) permite executar a ação de anotar com mais eficiência?
4. : Qual ou quais são as maneiras existentes de se reconhecer uma entidade dentro de um texto?

Protocolo da revisão sistemática de literatura	
Bases de Pesquisas	
	Acm - & http://dl.acm.org/
	IEEE Xplore - http://ieeexplore.ieee.org
	ScienceDirect - http://www.sciencedirect.com/
	Springer - http://link.springer.com/
Critério de Pesquisa	annotation and Linked Data.
	annotation and Linked Data or Linked Open Data
	semantic annotation and Linked Data
Critério de Seleção	
	Periódicos revisado por pares (Journals, Magazines and Conference Publication)
	Grande área: Ciências exatas da terra.
	Período: 2010 - 2014.
	Inglês e Português.
	Texto Completo.
Critério de Exclusão	
	Não disponibilizar o texto completo para leitura.
	Não estar relacionado com anotação semântica.
	Livros, Resenhas, Monografias.
	Estudos duplicados.
	Análise das palavras chave no abstract e título.
	Análise das palavras chave no conteúdo.
Categorização Definida	
	Base
	Tipo Produção
	Ano
	País
	Contribuição
	Ferramenta Anotacao
	Ferramenta de Extracao de Informação
	Extensível
	Plataforma
	Método de Reconhecimento da Entidade
	Forma de Anotação de Entidades
	Modo Salva as Anotações
	Formato Salva Anotações
	Formato Entrada Doc1

Figura 23 – Protocolo da Revisão Sistemática de Literatura

De acordo com figura 23, Protocolo da Revisão Sistemática de Literatura, ficou definido o protocolo de revisão sistemática da literatura definida (PRSL). O critério de pesquisa foi adotado nas bases de pesquisas afim de realizar de forma automatizada, com o auxilio do programa Zotero, a extração das produções observando os critérios de pesquisa e seleção. Como o núcleo do trabalho envolve anotação automática de documentos a partir de dados abertos, ficou definido que as palavras chave que serão utilizadas nas bases de pesquisas serão (annotation and Linked Data; annotation and Linked Data or Linked Open Data; semantic annotation and Linked Data). A idéia foi selecionar produções que abordem conceitos e práticas relacionadas a anotação semântica e Linked Open Data. As etapas de seleção das produções foi efetuada através da análise subjetiva, em seguida relevância das palavras chave no conteúdo, e por fim leitura das produções.

Nesta etapa de planejamento, o processo é iterativo e dinâmico enquanto a RSL estiver ocorrendo, os resultados são adaptados de acordo com os objetivos relacionados na revisão podendo eles serem delimitados de acordo com a evolução dos critérios.

5.4 Realização

Após o planejamento e definições mencionadas no PRSL, foi o momento de executar o processo de seleção e filtragem das publicações. Nessa etapa de realização da RSL é necessário identificação da pesquisa; seleção dos estudos primários de acordo com os critérios estabelecidos. Os critérios da pesquisa e seleção são ajustados para que se consiga alcançar um número de produções que seja relevante para o trabalho, dessa maneira, abaixo temos as seguintes consultas e resultados em cada base.

Construção da pesquisa na Base ACM:

Searching for: (Annotation) and ("linked data" or "linked open data") and (PublishedAs:journal OR PublishedAs:magazine)
Found **28** within *Publications from ACM and Affiliated Organizations* (Full-Text collection)

Figura 24 – String de pesquisa utilizada na base ACM

Searching for: (annotation) and ("linked open data"or "linked data") and (PublishedAs:journal OR PublishedAs:magazine) Sendo encontradas 28 produções.

Construção da pesquisa na IEEE Xplore:

Figura 25 – String de pesquisa utilizada na base IEEE

You searched for: ((annotation) AND "linked open data"OR "linked data") You Refined by Content Type: Conference Publications Remove , Journals Magazines Remove Publication Year: 2010 - 2014. Tendo como resultado 322 produções.

Construção da pesquisa na ScienceDirect:

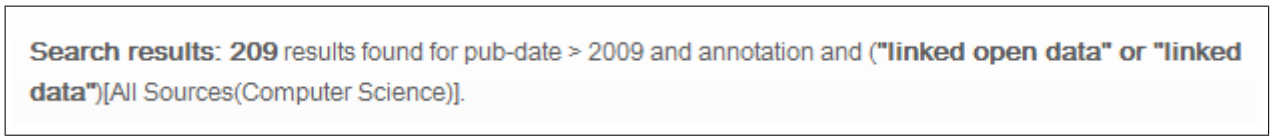

Figura 26 – String de pesquisa utilizada na base ScienceDirect

O resultado da pesquisa foi de: 209 produções encontradas para (annotation and ("linked data"or "linked open data")[All Sources(Computer Science)]).

Construção da pesquisa na Springer:

Figura 27 – String de pesquisa utilizada na base Springer

Foi localizado 240 produções com resultado para '(annotation and ("linked data"OR "linked open data"))' within Computer Science; Article; 2010 - 2015

Nessa etapa de seleção foi realizado uma análise das produções utilizando as palavras chave no abstract e título, através de um programa que foi criado na linguagem PHP com a finalidade de verificar se as palavras chave existem no abstract e título. O programa leu um arquivo «base».bib e retornou «base-excl01».bib. Obteve como resultado dessa primeira seleção:

ACM DE 28 PARA 24

IEE DE 321 PARA 318

SCIENCEDIRECT DE 229 PARA 192

SPRINGER DE 240 PARA 210

TOTAL DE: 818 PARA: 744

Nesta segunda e última etapa da análise das produções foi utilizado a ferramenta AlchemyAPI para gerar dados com relação a importância das palavras chaves dentro do contexto. Foi criado um programa em PHP para utilizar a API do Alkemy: Tem como entrada um arquivo «base-excl01».bib e retorno um arquivo «base-excl01Result».xls . Em realizou-se uma análise do resultado do arquivo (.xls) , observando e selecionando as publicações que continham palavras chave (annotation, semantic, linked Data, Linked Open Data) disponibilizada pela API, sempre observando a relevância desses termos dentro do documento.

ACM DE 24 PARA 3

IEE DE 318 PARA 13

SCIENCEDIRECT DE 192 PARA 11

SPRINGER DE 210 PARA 39

GOOGLE 15

TOTAL DE: 744 PARA: 81

5.5 Resultados

Esta seção apresentará os resultados encontrados na RSL, sendo a base do trabalho dessa pesquisa e condizente com o planejamento proposto. A leitura e análise dos artigos propostos na seção anterior, Realização, estão ocorrendo, sendo apresentado abaixo uma pré analise do que foi identificado até o momento e pode vim a responder as questões levantadas na seção de Planejamento.

Com relação a Questão 1: Qual ou quais componentes/módulos que são necessários para montar uma arquitetura de anotação semântica com Linked Open Data?

As leituras realizadas até o presente momento, onde podemos citar os autores

(FAFALIOS; PAPADAKOS, 2014), (NETO, 2009), (FONTES; CAVALCANTI; MOURA, 2013) indicam uma padrão quanto ao modelo de framework de anotação semântica: um objeto responsável pela análise e extração dos termos de um documento e um objeto responsável pela criação de um documento anotado. A variação encontra-se principalmente no quesito da base de dados utilizada para efetuar o mapeamento dos termos encontrados onde no caso (NETO, 2009), (FONTES; CAVALCANTI; MOURA, 2013) utiliza-se ontologias específicas para o desenvolvimento dos seus trabalhos e no caso do (FAFALIOS; PAPADAKOS, 2014) utiliza as ontologias disponíveis no LOD.

Questão 2: Dos componentes/módulos encontrados quais são os mais utilizados?

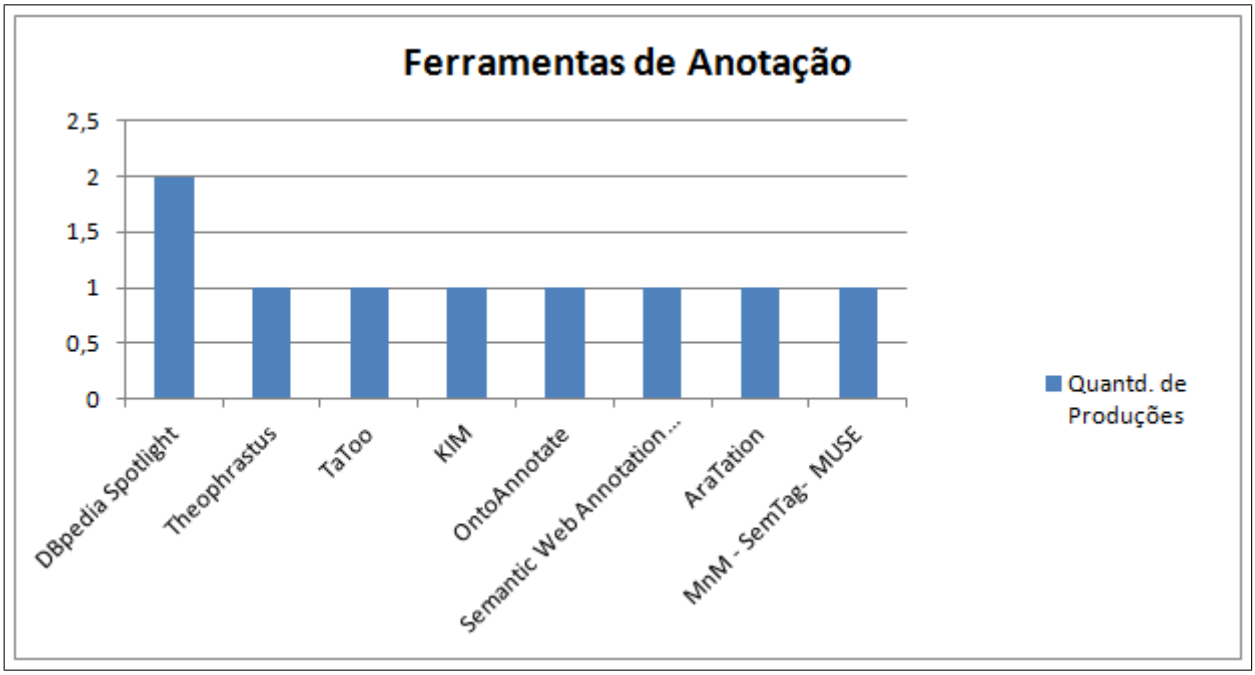

Figura 28 – Quantidade de Ferramentas de Anotação Identificada na RSL

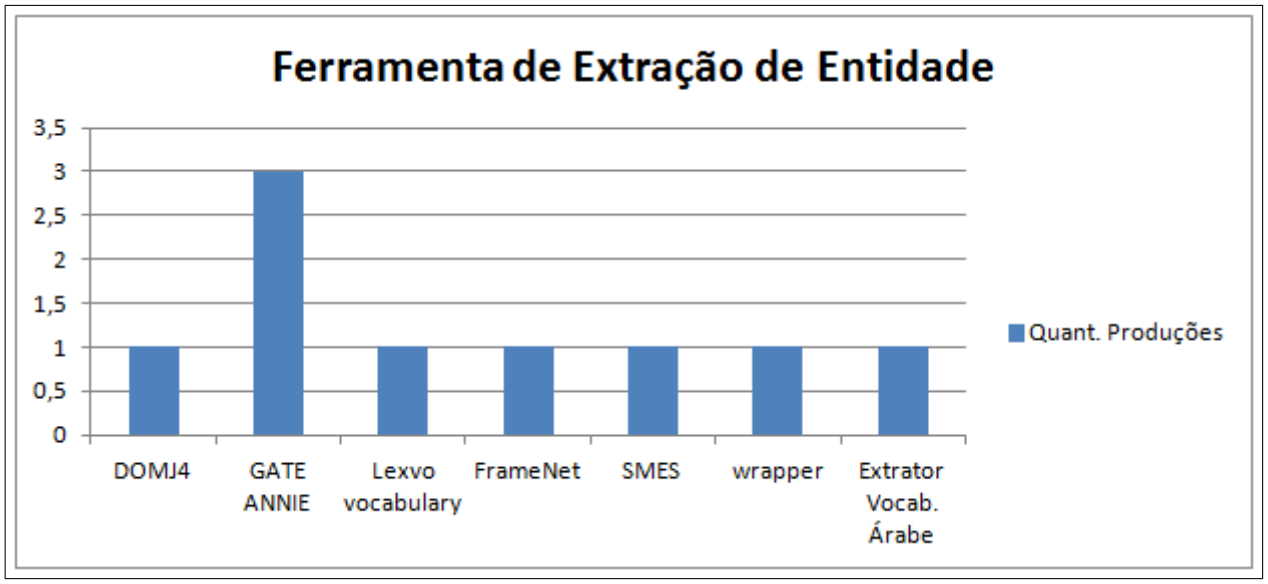

Figura 29 – Quantidade de Ferramentas de Extração de Entidade Identificada na RSL

Questão 3: Dos componentes/módulos encontrados, qual(is) permite executar a ação de anotar com mais eficiência?

Na pesquisa realizada por (REEVE; HAN, 2005) ele comparou algumas ferramentas para anotação semântica disponíveis até o presente ano da sua pesquisa, ressaltou suas características e apurou a eficácia de suas anotações. Ele destacou que a ferramenta MnM, que utiliza aprendizagem de máquina na identificação das entidades, como a de melhor desempenho e a de pior a Onto-O-Mat. Em suas conclusões ele informa que algoritmos de aprendizagem de máquina são mais efetivos do que os métodos baseados em padrões, porém os sistemas baseados em regras podem possuir uma performance melhor do que os sistemas baseados em aprendizagem de máquina.

Questão 4: Qual ou quais são as maneiras existentes de se reconhecer uma entidade dentro de um texto?

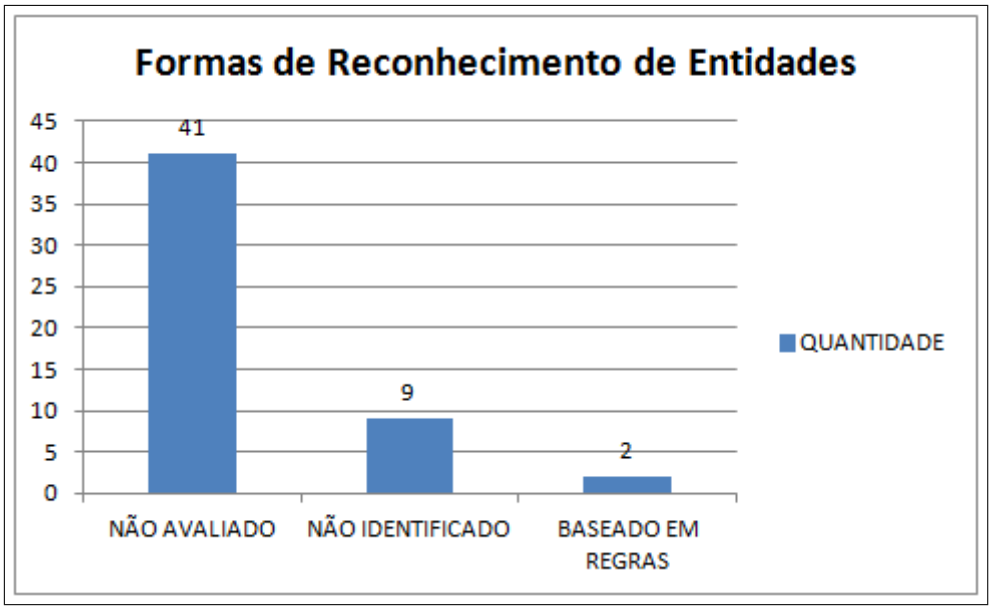

Figura 30 – Formas Encontradas de Reconhecer uma Entidade

6 Metodologia

Conforme (GIL, 2002), a pesquisa experimental consiste em determinar um objeto de estudo, selecionar as variáveis que seriam capazes de influenciá-lo, definir as formas de controle e observação dos efeitos que a variável produz no objeto. Ela caracteriza-se por manusear sem desvios as variáveis associadas com o núcleo de estudo. De acordo com (CERVO; BERVIAN; SILVA, 2006), a experimentação é um conjunto de processos usados para conferir as hipóteses, sendo essa uma relação de causa e efeito ou de antecedência e consequência entre os dois acontecimentos.

A natureza da pesquisa será aplicada, porque haverá um interesse em adquirir novos conhecimentos, porém orientada para a aplicação prática. Essa pesquisa é realizada para determinar as possíveis utilidades para as descobertas da pesquisa básica ou definir novos métodos ou definir novas maneiras de alcançar a solução de problemas específicos (CASARIN; CASARIN, 2011). Apesar da natureza da pesquisa ser do tipo aplicada, ligada a prática, essa não pode deixar de incluir consideração e pensamentos teóricos(MASCARENHAS,).

Com relação ao problema a ser estudado, esse é qualitativo, pois prevalece a parte descritiva, onde os objetivos envolvem a descrição de certo fenômeno, caracterizando sua ocorrência e relacionando-o com outros fatores (CASARIN; CASARIN, 2011). Segundo (MASCARENHAS,) a pesquisa qualitativa é feita quando deseja-se descrever com mais detalhes um determinado objeto de estudo. Neste âmbito, a pesquisa será explicativa, pois procura identificar fatores que determinam ou contribuem para a ocorrência dos fenômenos (GIL, 2002).

Assim por intermediação das definições de tipologias de pesquisas apresentadas, pode-se afirmar que a pesquisa do tipo prova de conceito e experimento é um caminho adequado para o desenvolvimento dessa pesquisa. Em pesquisas explicativas o método utilizado é o experimental, sendo utilizado principalmente na área de ciências exatas (CASARIN; CASARIN, 2011).

Atividades Metológicas	**OBJETIVOS ESPECÍFICOS**		
	Conceituar e identificar as tecnologias relacionadas com Anotação Semântica.	Selecionar e Extrair currículos dos docentes da plataforma Lattes.	Anotar o Lattes.
Revisão Sistemática da Literatura			
Seleção dos nomes de docentes selecionados na CAPES dos documentos PDFs.			
A partir dos nomes selecionados, o DCC/UFMG, importará os currículos Lattes dos docentes.			
Identificação dos DataSets disponíveis e a forma de interligá-los.			
Identificação e seleção dos softwares necessários para anotar do Lattes.			
Arcabouço de anotação do Lattes.			

Figura 31 – Relação entre: Atividades Metodológicas X Objetivos Específicos

Os procedimentos metodológicos para atingir o objetivos específicos estão abordados no figura31. O trabalho será desenvolvido em duas partes, a primeira constará do levantamento bibliográfico sobre Anotação Semântica abrangendo conceitos, componentes, ferramentas, tecnologias e suas aplicações, com o objetivo de analisar de que maneira ela pode ser utilizada no desenvolvimento do arcabouço. Na segunda parte, será criado de acordo com o arcabouço a aplicabilidade dos conceitos, ferramentas e tecnologias para a execução da Anotação Semântica.

7 Arcabouço Conceitual

O objetivo deste trabalho será anotar automaticamente os documentos Web do Currículo Lattes utilizando as ontologias disponíveis e as tecnologias de Anotação Semântica.

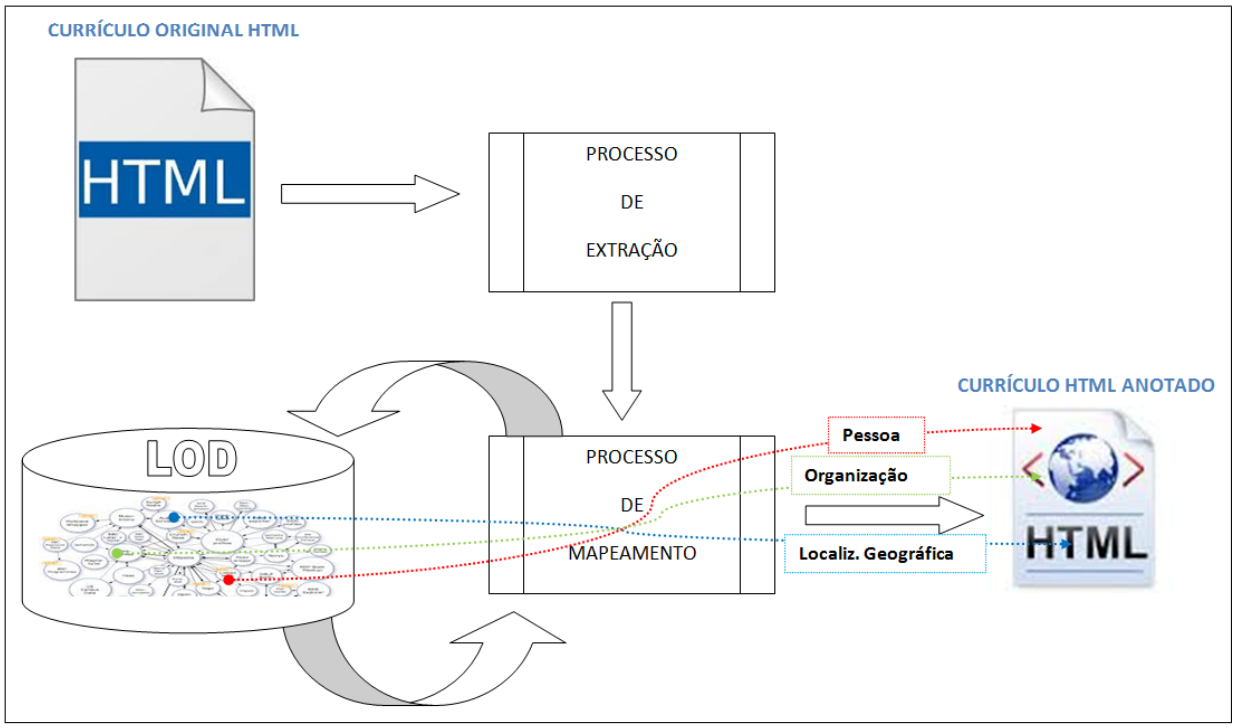

Figura 32 – Modelo conceitual do projeto

A figura 32 representa a proposta da arquitetura conceitual de anotação semântica e tem como objetivo identificar os principais componentes e etapas que farão parte desta dissertação. Podemos explicar o processo nas seguintes etapas:

1. É passado para o sistema a URL de um determinado currículo lates.
2. O Processo de Extração tem a função de extrair os termos do currículo, em seguida passando para o Processo de Mapeamento.
3. O Mapeamento é realizado de maneira que possibilite identificar entidades de um domínio de interesse a partir dos termos identificados no currículo.
4. No fim do processo, com a identificação da entidade que permite disponibilizar a semântica a um termo do documento, será criado um novo documento no formato RDFa.

O novo documento anotado com os dados semânticos, estará legível tanto para as pessoas quanto para as máquinas, abrindo possibilidades para os motores de busca inteligentes. Para a execução deste modelo será realizado um levantamento na RSL de quais ferramentas estão disponíveis para executar essa tarefa utilizando o conjunto de informações disponíveis no LOD.

Referências

BELLOZE, K. T. et al. An evaluation of annotation tools for biomedical texts. In: CITESEER. *ONTOBRAS-MOST*. 2012. p. 108–119. Disponível em: <http://ceur-ws.org/Vol-938/ontobras-most2012_paper9.pdf>.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web:a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, v. 284, n. 5, p. 34–43, may 2001. Disponível em: <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>.

BIZER TOM HEATH, T. B. C. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, v. 5, n. 3, p. 1–22, 2009. Disponível em: <http://dx.doi.org/10.4018/jswis.2009081901>.

BONIFACIO, A. S. *Ontologias e consulta semântica : uma aplicação ao caso Lattes*. Dissertação (Mestrado) — UFRGS, Porto Alegre, 2002. Disponível em: <http://www.lume.ufrgs.br/handle/10183/7082>.

CASARIN, H. d. C. S.; CASARIN, S. J. C. *Pesquisa Científica: da teoria à prática*. [S.l.: s.n.], 2011.

CASTAñO, A. C. *Populando ontologias através de informações em HTML - o caso do currículo lattes*. Dissertação (Mestrado) — Universidade de São Paulo, 2008. Disponível em: <http://www.teses.usp.br/teses/disponiveis/45/45134/tde-12082008-130204/>.

CERVO, A.; BERVIAN, P.; SILVA, R. da. *Metodologia científica*. Pearson Prentice Hall, 2006. ISBN 9788576050476. Disponível em: <https://books.google.com.br/books?id=9SK2GQAACAAJ>.

CNPQ. Plataforma lattes cnpq. In: . [s.n.], 2014. Disponível em: <http://lattes.cnpq.br/>.

ELLER, M. P. *Anotações Semânticas de Fontes de Dados Heterogêneas Um Estudo de Caso com a Ferramenta Smore*. Dissertação (Mestrado) — Dissertação de Mestrado–UFSC–Departamento de Informática e Estatística, 2008.

FAFALIOS, P.; PAPADAKOS, P. Theophrastus: On demand and real-time automatic annotation and exploration of (web) documents using open linked data. *Web Semantics: Science, Services and Agents on the World Wide Web*, n. 0, p. –, 2014. ISSN 1570-8268. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1570826814000572>.

FONTES, C. A.; CAVALCANTI, M.; MOURA, A. D. C. An ontology-based reasoning approach for document annotation. p. 160–167, Sept 2013. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6693512>.

FONTES, C. A.; MOURA, A. M. de C.; CAVALCANTI, M. C. Anotação semântica em documentos. In: . [s.n.], 2010. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/wtdbd/2010/sbbd_wtd_14.pdf>.

FONTES, C. A. et al. Recuperação de informações em documentos anotados semanticamente na Área de gestão ambiental. p. 43–52, 2010. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/ontobras/2010/004.pdf>.

GALEGO, E. F. Extração e consulta de informações do currículo lattes baseada em ontologias. 2013. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/eniac/2013/0043.pdf>.

GIL, A. C. *Como elaborar projetos de pesquisa*. [S.l.: s.n.], 2002.

MASCARENHAS, S. *METODOLOGIA CIENTIFICA*. PEARSON BRASIL. ISBN 9788564574595. Disponível em: <https://books.google.com.br/books?id=kOZBLgEACAAJ>.

NETO, G. M. d. S. *Anotacao Semantica De Recursos Web Baseada em Ontologias*. Dissertação (Mestrado) — Dissertação de Mestrado–UFAM–INSTITUTO DE CIÊNCIAS EXATAS–PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA, 2009. Disponível em: <http://www.dominiopublico.gov.br/pesquisa/DetalheObraForm.do?select_action=&co_obra=148061>.

OREN, E. et al. What are semantic annotations. In: . [S.l.]: Citeseer, 2006.

PLANETDATA. Linked open data cloud diagram 2014. In: . [s.n.], 2014. Disponível em: <http://data.dws.informatik.uni-mannheim.de/lodcloud/2014/>.

REEVE, L.; HAN, H. Survey of semantic annotation platforms. In: *Proceedings of the 2005 ACM Symposium on Applied Computing*. New York, NY, USA: ACM, 2005. (SAC '05), p. 1634–1638. ISBN 1-58113-964-0. Disponível em: <http://doi.acm.org/10.1145/1066677.1067049>.

VIRGILIO, R. D. et al. A reverse engineering approach for automatic annotation of web pages. *Multimedia Tools and Applications*, v. 64, n. 1, p. 119–140, may 2013. ISSN 1380-7501, 1573-7721. 00001. Disponível em: <http://link.springer.com/article/10.1007/s11042-011-0852-8>.

W3C. Owl web ontology language. In: . [s.n.], 2014. Disponível em: <http://www.w3.org/TR/owl-features/>.

W3C. Owl2 - web ontology language 2. In: . [s.n.], 2015. Disponível em: <http://www.w3.org/TR/2012/REC-owl2-overview-20121211/>.

W3C$_{D}ADOS_{A}BERTOS.Dadosabertosgovernamentais.In$: .[$s.n.$], 2014.$Disponívelem$: <>.

W3C$_{R}DF1.1_{P}RIMER.Rdf1.1primer.In$: .[$s.n.$], 2014.$Disponívelem$: <>.

W3C$_{S}CHEMA1.1.Rdfschema1.1.In$: .[$s.n.$], 2014.$Disponívelem$: <>.

ZHANG, Z.; CHEN, S.; FENG, Z. Semantic annotation for web services based on DBpedia. p. 280–285, 2013.